

工程设计与分析系列

SPSS 统计分析与数据挖掘 (第3版)

谢龙汉 蔡思祺 编著

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书基于 SPSS 24.0 编写,是在修正并完善第 2 版的基础上完成的;每章均有大量分析案例,结合案例对 SPSS 各模块的统计分析功能和图形功能进行详细讲解。本书具体内容为 SPSS 简介、SPSS 数据挖掘系统介绍、数据文件管理、数据预处理、基本统计分析、多重反应分析、均值的比较与检验、统计图制作、参数检验、回归分析、方差分析、相关分析、聚类分析、判别分析、因子分析、对应分析、信度分析、生存分析、对数线性模型、时间序列分析、缺失值分析,以及 SPSS 在财务智能、数据预测、股市分析、社会经济分析、金融数据分析等方面的数据挖掘应用。

本书最大特点是抛弃了其他同类书籍中只介绍理论用法、缺乏案例分析的弊端,全书给出大量数据挖掘分析案例,并配有视频讲解,为读者展示 SPSS 在数据分析、信用风险管理、直销分析、社会经济分析等实际项目中的应用。

本书适合众多领域的数据分析人员,也可供相关专业本科生、研究生、科技人员和企事业单位工作人员,以及从事数据挖掘、金融分析、商业咨询、财务分析的人员学习。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有,侵权必究。

图书在版编目(CIP)数据

SPSS 统计分析与数据挖掘 / 谢龙汉, 蔡思祺编著. —3 版. —北京: 电子工业出版社, 2017.11

(工程设计与分析系列)

ISBN 978-7-121-32907-4

I. ①S… II. ①谢… ②蔡… III. ①统计分析—统计程序 IV. ①C819

中国版本图书馆 CIP 数据核字(2017)第 258342 号

策划编辑: 许存权

责任编辑: 许存权 特约编辑: 谢忠玉 等

印 刷: 三河市良远印务有限公司

装 订: 三河市良远印务有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本: 787×1 092 1/16 印张: 31.75 字数: 812 千字

版 次: 2012 年 1 月第 1 版

2017 年 11 月第 3 版

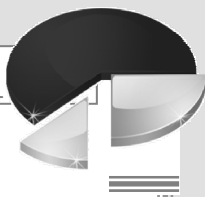
印 次: 2017 年 11 月第 1 次印刷

定 价: 79.00 元(含光盘 1 张)

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010) 88254888, 88258888。

质量投诉请发邮件至 zlts@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式:(010) 88254484, xucq@phei.com.cn。



再版前言

SPSS (Statistical Product and Service Solutions, 统计产品和服务解决方案) 是当今国际上最流行的统计分析软件之一, 具有界面友好、统计功能强大、前后处理功能完善等优点。SPSS 提供了广泛的数据收集、分类、分析和处理技术, 揭示了数据模式、异常, 以及关键变量和关系, 帮助企业深入洞察企业信息, 做出更好决策。本书从 SPSS 窗口操作出发, 用案例的形式介绍 SPSS 数据分析在各个领域的广泛应用。


本书集作者多年使用 SPSS 的工作经验, 并在改正第 2 版错误的基础上编写, 在编写过程中, 突出了以下特点。

直观易懂性。全书以图解实例的形式介绍基础知识和实例操作, 所有的知识模块和案例分析都尽可能详细, 模块操作采取中英文介绍的方式进行, 直观易懂, 使读者能够在最短的时间内获取最多的知识。

先进性。以最新的 SPSS 24.0 中文版为蓝本进行讲解, 中英文并用, 广泛吸收国内外优秀教材的成果进行内容编排, 在系统介绍基本理论和基本方法的同时, 注意介绍新的成熟的内容, 以及统计学在实际问题中的应用。

实用性。全书采用了基础知识介绍和实例操作相结合的方法, 互相补充, 书中的实例大多来源于经济生活之中, 使读者在学完本书后能够快速将知识应用于实践。

结构清晰, 讲解详尽。全书采用基础知识—窗口操作—综合实例分析的循序渐进的讲解方法, 一步一步地提高读者的 SPSS 操作知识, 而且每个知识点和实例都尽可能详细地讲解, 使读者学习起来轻松自如。

全部的案例数据、程序与多媒体示范相结合。本书的配套光盘  中提供了所有实例的数据、SPSS 操作视频, 读者可以在观看录像中增强对知识点的理解。

本书共 24 章, 依次介绍 SPSS 基本文件管理、基本统计分析、高级统计分析、决策树模型、神经网络模型、信用风险、社会经济评价, 以及各章节中的案例分析等内容。

第 1 章 SPSS 软件概述。包括 SPSS 软件简介、SPSS 操作入门、SPSS 各个模块, 以及 SPSS 帮助系统。

第 2 章 SPSS 数据挖掘系统。包括数据挖掘概述、SPSS 数据挖掘过程的介绍, 以便掌握数据挖掘基本概念、流程等知识。

第 3 章 数据文件、变量与函数。包括 SPSS 的变量类型、SPSS 数据文件的打开和保存, 最后介绍 SPSS 的函数。

第 4 章 数据预处理。包括最基本的数据文件的整理和数据变量的变换和计算。

第 5 章 基本统计分析。包括基本概念、频数过程、描述性统计分析过程、数据探索性分析过程, 以及交叉表分析过程。

第 6 章 参数检验。包括参数估计和假设检验的概述、平均值过程、单样本 t 检验、独立样本 t 检验及成对样本 t 检验。

第 7 章 基本图形的绘制。包括统计图概述、条形图、折线图、面积图、饼图、高低图、质量控制图、箱图、散点图、直方图、P-P 图和 Q-Q 图, 以及时间序列图。

第8章 非参数检验。包括非参数检验概述、 χ^2 检验、二项分布检验、游程检验、K-S检验、两独立样本分布位置检验、多个独立样本分布位置检验、两个相关样本分布位置检验、多个相关样本分布位置检验。

第9章 方差分析。包括方差分析的基本原理、单因素方差分析、多因素方差分析和协方差分析。

第10章 回归分析。包括线性回归、非线性回归,以及 Logistic 回归过程。

第11章 相关分析。包括相关分析概述、双变量相关过程、偏相关分析过程,以及距离过程。

第12章 聚类分析。包括聚类分析的原理、快速聚类的分析过程、系统聚类的分析过程、二阶聚类的分析过程,以及实例分析。

第13章 判别分析。包括判别分析的基本原理、一般判别分析过程和逐步判别分析过程。

第14章 因子分析。包括因子分析概述及 SPSS 中因子分析的操作过程。

第15章 对应分析。包括对应分析的基本原理、对应分析过程及最优标度分析过程。

第16章 可靠性和多维尺度分析。包括可靠性和多维标度的概述、分析过程及实例。

第17章 生存分析。包括生存分析概述、寿命表分析过程、Kaplan-Meier 分析过程及 Cox 模型回归分析过程。

第18章 对数线性模型。包括对数线性模型概述、常规模型分析过程、分对数分析过程及选择模型分析过程。

第19章 时间序列分析。包括时间序列概述、时间序列数据的预处理、指数平滑方法、ARIMA 模型及季节性分解模型分析过程。

第20章 缺失值分析。包括 SPSS 中的缺失值理论概述、SPSS 缺失值分析的操作过程,以及缺失值实例分析。

第21章 决策树模型。包括决策树模型概述、SPSS 中决策树的参数设置,以及利用实例分析来介绍决策树模型的应用过程。

第22章 神经网络。包括神经网络概述、神经网络模型分析参数的设置及实例分析。

第23章 信用风险分析。包括主要信用风险概述,以及利用 SPSS 解决信用风险的各种实例分析。

第24章 SPSS 在社会经济综合评价中的应用。包括 SPSS 的各种分析案例,包括沿海省市经济综合指标的主成分分析、中国城镇居民消费结构的聚类分析研究,以及我国内地可支配收入和消费性支出之间的回归分析。

本书主要由谢龙汉、蔡思祺完成,参与编著和光盘开发的还有林伟、魏艳光、林木议、王悦阳、林伟洁、林树财、郑晓、吴苗、李翔、朱小远、唐培培、耿煜、邓奕、张桂东、鲁力、于斌、尚涛、黄海等。由于时间仓促,书中难免有疏漏之处,请读者谅解。读者可通过电子邮件 xielonghan@aliyun.com.cn 与我们交流。

注:本书在介绍软件应用时,命令、选项等包含英文注释,有助于使用英文版软件的读者学习。

编著者



目 录

| | | | |
|--------------------------|----|--|----|
| 第 1 章 SPSS 软件概述 | 1 | 3.2.3 数据文件保存 | 38 |
| 1.1 SPSS 简介 | 1 | 3.3 SPSS 函数 | 38 |
| 1.2 SPSS 操作入门 | 2 | 3.3.1 算术函数 | 39 |
| 1.2.1 软件安装、启动及退出 | 3 | 3.3.2 统计函数 | 39 |
| 1.2.2 操作环境 | 4 | 3.3.3 逻辑函数 | 40 |
| 1.2.3 系统参数的设置 | 7 | 3.3.4 日期和时间函数 | 40 |
| 1.3 SPSS 的帮助系统 | 15 | 3.3.5 随机变量函数 | 42 |
| 第 2 章 SPSS 数据挖掘系统 | 17 | 3.3.6 反分布函数 | 43 |
| 2.1 数据挖掘概述 | 17 | 3.3.7 累计分布函数 | 44 |
| 2.1.1 数据挖掘的含义 | 17 | 3.3.8 缺失值函数 | 46 |
| 2.1.2 数据挖掘与 OLAP | 18 | 3.3.9 字符串函数 | 47 |
| 2.1.3 数据挖掘和统计学 | 18 | 第 4 章 数据预处理 | 49 |
| 2.1.4 数据挖掘的目的 | 19 | 4.1 数据文件的整理 | 49 |
| 2.1.5 数据挖掘应用 | 19 | 4.1.1 个案排序 (Sort Case) 过程 | 50 |
| 2.1.6 数据挖掘流程 | 19 | 4.1.2 转置 (Transpose) 过程 | 50 |
| 2.2 成功的数据挖掘 | 20 | 4.1.3 合并文件 (Merge File) 过程 | 51 |
| 2.2.1 CRISP-DM 方法论 | 21 | 4.1.4 汇总 (Aggregate) 过程 | 53 |
| 2.2.2 选择数据挖掘工具 | 25 | 4.1.5 拆分文件 (Split File) 过程 | 55 |
| 2.2.3 SPSS 数据挖掘 | 26 | 4.1.6 选择个案 (Select Cases) 过程 | 55 |
| 2.3 SPSS 数据挖掘的过程 | 29 | 4.1.7 个案加权 (Weight Cases) 过程 | 56 |
| 2.3.1 商业理解 | 29 | 4.2 数据变量的变换和计算 | 56 |
| 2.3.2 数据理解 | 29 | 4.2.1 计算变量 (Compute Variables) 过程 | 57 |
| 2.3.3 数据准备 | 29 | 4.2.2 计数 (Count) 过程 | 59 |
| 2.3.4 数据模型 | 30 | 4.2.3 重新编码 (Recode) 过程 | 60 |
| 2.3.5 评估 | 30 | 4.2.4 个案排秩 (Rank Cases) 过程 | 61 |
| 2.3.6 部署 | 31 | 4.2.5 自动重新编码 (Automatic Recode) 过程 | 63 |
| 第 3 章 数据文件、变量与函数 | 33 | | |
| 3.1 SPSS 的变量类型 | 33 | | |
| 3.1.1 数据的输入 | 34 | | |
| 3.1.2 变量的编辑 | 35 | | |
| 3.2 数据文件的打开和保存 | 36 | | |
| 3.2.1 打开 SPSS 数据文件 | 37 | | |
| 3.2.2 打开其他格式的数据文件 | 37 | | |

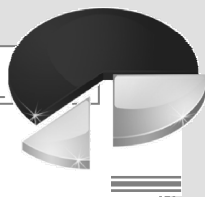
| | | | |
|--------------------------|-----|------------------------|-----|
| 第5章 基本统计分析 | 65 | 第7章 基本图形的绘制 | 103 |
| 5.1 基本概念 | 65 | 7.1 统计图概述 | 103 |
| 5.1.1 基本的统计概念 | 65 | 7.2 条形图 | 104 |
| 5.1.2 描述性统计分析 | 67 | 7.3 折线图 | 108 |
| 5.2 频率分析 | 68 | 7.4 面积图 | 110 |
| 5.2.1 频率分析过程的操作界面 | 68 | 7.5 饼图 | 111 |
| 5.2.2 实例分析 | 70 | 7.5.1 饼图参数设置 | 111 |
| 5.3 描述性统计分析过程 | 72 | 7.5.2 实例分析 | 112 |
| 5.3.1 描述性统计分析过程参数设置 | 72 | 7.6 高低图 | 113 |
| 5.3.2 实例分析 | 72 | 7.7 质量控制图 | 114 |
| 5.4 数据探索性分析过程 | 74 | 7.8 箱图 | 119 |
| 5.4.1 数据探索性分析过程参数设置 | 74 | 7.8.1 箱图参数设置 | 119 |
| 5.4.2 实例分析 | 75 | 7.8.2 实例分析 | 120 |
| 5.5 交叉表分析过程 | 78 | 7.9 散点图 | 121 |
| 5.5.1 交叉表过程的参数设置 | 78 | 7.9.1 散点图参数设置 | 122 |
| 5.5.2 实例分析 | 81 | 7.9.2 实例分析 | 122 |
| 第6章 参数检验 | 84 | 7.10 直方图 | 124 |
| 6.1 参数估计和假设检验概述 | 84 | 7.11 P-P图和Q-Q图 | 124 |
| 6.1.1 参数估计 | 84 | 7.12 时间序列图 | 126 |
| 6.1.2 假设检验 | 87 | 7.12.1 时间序列图参数设置 | 126 |
| 6.2 平均值(Means)过程 | 92 | 7.12.2 实例分析 | 130 |
| 6.2.1 SPSS的平均值过程参数的设置 | 92 | 第8章 非参数检验 | 133 |
| 6.2.2 平均值过程实例 | 93 | 8.1 非参数检验概述 | 133 |
| 6.3 单样本 t 检验 | 94 | 8.2 χ^2 检验 | 134 |
| 6.3.1 单样本 t 检验过程的参数设置 | 94 | 8.2.1 χ^2 检验的参数设置 | 135 |
| 6.3.2 实例分析 | 95 | 8.2.2 χ^2 检验实例分析 | 137 |
| 6.4 独立样本 t 检验 | 97 | 8.3 二项分布检验 | 139 |
| 6.4.1 独立样本 t 检验过程的参数设置 | 97 | 8.3.1 二项分布检验的参数设置 | 139 |
| 6.4.2 实例分析 | 98 | 8.3.2 实例分析 | 139 |
| 6.5 成对样本 t 检验 | 100 | 8.4 游程检验 | 141 |
| 6.5.1 成对样本 t 检验过程的参数设置 | 100 | 8.4.1 游程检验的参数设置 | 142 |
| 6.5.2 实例分析 | 100 | 8.4.2 实例分析 | 142 |
| | | 8.5 单样本K-S检验 | 144 |
| | | 8.5.1 单样本K-S检验的参数设置 | 144 |
| | | 8.5.2 实例分析 | 145 |

| | | | |
|-------------------------------------|-----|--|-----|
| 8.6 两个独立样本分布位置检验 | 147 | 第 10 章 回归分析 | 187 |
| 8.6.1 两个独立样本分布位置检验的 参数设置 | 148 | 10.1 线性回归 | 187 |
| 8.6.2 实例分析 | 148 | 10.1.1 线性回归模型 | 188 |
| 8.7 多个独立样本分布位置检验 | 150 | 10.1.2 最小二乘估计 | 188 |
| 8.7.1 多个独立样本分布位置检验的 参数设置 | 150 | 10.1.3 回归方程的显著性检验 | 189 |
| 8.7.2 实例分析 | 151 | 10.1.4 预测问题 | 191 |
| 8.8 两个相关样本分布位置检验 | 153 | 10.1.5 SPSS 线性回归分析设置 | 192 |
| 8.8.1 两个相关样本分布位置检验的 参数设置 | 153 | 10.1.6 回归分析模型的实例分析 | 196 |
| 8.8.2 实例分析 | 154 | 10.2 非线性回归 | 199 |
| 8.9 多个相关样本分布位置检验 | 155 | 10.2.1 非线性回归分析的基本原理 | 200 |
| 8.9.1 多个相关样本分布位置检验的 参数设置 | 156 | 10.2.2 非线性回归参数设置 | 200 |
| 8.9.2 实例分析 | 156 | 10.2.3 实例分析 | 203 |
| 第 9 章 方差分析 | 159 | 10.3 Logistic 回归 | 205 |
| 9.1 方差分析的基本原理 | 159 | 10.3.1 Logistic 回归模型概述 | 206 |
| 9.1.1 自由度与平方和分解 | 160 | 10.3.2 二元 Logistic 回归模型参数 设置 | 207 |
| 9.1.2 F 检验 | 162 | 10.3.3 实例分析 | 210 |
| 9.1.3 多重比较 | 163 | 第 11 章 相关分析 | 215 |
| 9.2 单因素 ANOVA 检验 | 164 | 11.1 相关分析概述 | 215 |
| 9.2.1 单因素 ANOVA 检验步骤 | 165 | 11.1.1 相关关系 | 215 |
| 9.2.2 判断与结论 | 166 | 11.1.2 相关图形和相关系数 | 216 |
| 9.2.3 单因素 ANOVA 检验过程的 参数设置 | 167 | 11.1.3 SPSS 的相关分析功能简介 | 218 |
| 9.2.4 实例分析 | 169 | 11.2 双变量 (Bivariate) 过程 | 218 |
| 9.3 多因素方差分析 | 170 | 11.2.1 双变量相关分析简介 | 218 |
| 9.3.1 只考虑主效应的多因素方差 分析 | 171 | 11.2.2 双变量过程的参数设置 | 220 |
| 9.3.2 存在交互效应的多因素方差 分析 | 173 | 11.2.3 实例分析 | 222 |
| 9.3.3 单变量过程参数设置 | 175 | 11.3 偏相关 (Partial) 过程 | 224 |
| 9.3.4 实例分析 | 179 | 11.3.1 偏相关过程的参数设置 | 224 |
| 9.4 协方差分析 | 183 | 11.3.2 实例分析 | 225 |
| 9.4.1 协方差分析概述 | 183 | 11.4 Distances (距离) 过程 | 227 |
| 9.4.2 实例分析 | 184 | 11.4.1 Distances 过程的距离分析 参数设置 | 227 |
| | | 11.4.2 实例分析 | 230 |

| | | | |
|-----------------------------|-----|-------------------------------------|-----|
| 第 12 章 聚类分析 | 232 | 14.2.2 实例分析 | 286 |
| 12.1 聚类分析的原理 | 232 | 第 15 章 对应分析 | 291 |
| 12.1.1 一般原理 | 233 | 15.1 对应分析的基本原理 | 291 |
| 12.1.2 聚类分析步骤 | 236 | 15.2 对应分析 | 293 |
| 12.1.3 系统聚类方法 | 237 | 15.2.1 对应分析过程的参数设置 | 293 |
| 12.2 快速样本聚类过程 | 240 | 15.2.2 实例分析 | 296 |
| 12.2.1 快速聚类简介 | 240 | 15.3 最优标度过程 | 299 |
| 12.2.2 SPSS 快速聚类的设置 | 240 | 15.3.1 最优标度过程的参数设置 | 299 |
| 12.2.3 实例分析 | 242 | 15.3.2 实例分析 | 306 |
| 12.3 系统聚类过程 | 246 | 第 16 章 可靠性和多维标度分析 | 310 |
| 12.3.1 系统聚类简介 | 246 | 16.1 可靠性分析 | 310 |
| 12.3.2 SPSS 系统聚类设置 | 246 | 16.1.1 可靠性分析的基本原理 | 310 |
| 12.3.3 实例分析 | 249 | 16.1.2 可靠性分析的参数设置 | 312 |
| 12.4 二阶聚类分析 | 252 | 16.1.3 实例分析 | 314 |
| 12.4.1 二阶聚类简介 | 252 | 16.2 多维标度分析 | 316 |
| 12.4.2 SPSS 二阶聚类的设置 | 253 | 16.2.1 多维标度分析简介 | 316 |
| 12.4.3 实例分析 | 254 | 16.2.2 多维标度过程的参数设置 | 317 |
| 第 13 章 判别分析 | 257 | 16.2.3 实例分析 | 320 |
| 13.1 判别分析的基本原理 | 257 | 第 17 章 生存分析 | 323 |
| 13.1.1 判别分析简介 | 257 | 17.1 生存分析简介 | 323 |
| 13.1.2 判别分析的数学模型与判别方法 | 258 | 17.1.1 生存分析的基本概念 | 323 |
| 13.2 一般判别分析 | 265 | 17.1.2 生存资料的特点 | 325 |
| 13.2.1 一般判别分析的参数设置 | 265 | 17.1.3 生存分析方法 | 326 |
| 13.2.2 实例分析 | 267 | 17.1.4 SPSS 中的生存分析过程 | 326 |
| 13.3 逐步判别分析 | 272 | 17.2 寿命表 (Life Tables) 过程 | 327 |
| 13.3.1 逐步判别的参数设置 | 272 | 17.2.1 寿命表分析过程的参数设置 | 327 |
| 13.3.2 实例分析 | 273 | 17.2.2 实例分析 | 328 |
| 第 14 章 因子分析 | 279 | 17.3 Kaplan-Meier 分析 | 332 |
| 14.1 因子分析简介 | 279 | 17.3.1 Kaplan-Meier 分析过程的参数设置 | 332 |
| 14.1.1 因子分析的基本原理 | 280 | 17.3.2 实例分析 | 334 |
| 14.1.2 因子分析的基本步骤和过程 | 282 | 17.4 Cox 模型回归分析 | 337 |
| 14.2 SPSS 因子分析 | 283 | | |
| 14.2.1 SPSS 因子分析的参数设置 | 283 | | |

| | | | |
|--------------------------------------|------------|----------------------------------|------------|
| 17.4.1 Cox 回归模型 | 337 | 19.4.1 ARIMA 模型的基本原理 | 386 |
| 17.4.2 Cox 模型分析过程的参数设置 | 339 | 19.4.2 ARIMA 模型分析过程的参数设置 | 389 |
| 17.4.3 实例分析 | 343 | 19.4.3 实例分析 | 390 |
| 第 18 章 对数线性模型 | 348 | 19.5 季节性分解模型 | 394 |
| 18.1 对数线性模型概述 | 348 | 19.5.1 季节性分解模型分析过程的参数设置 | 394 |
| 18.2 常规模型 (General) 过程 | 349 | 19.5.2 实例分析 | 395 |
| 18.2.1 常规模型分析过程的参数设置 | 349 | 第 20 章 缺失值分析 | 399 |
| 18.2.2 实例分析 | 351 | 20.1 缺失值理论概述 | 399 |
| 18.3 分对数 (Logit) 过程 | 354 | 20.1.1 数据缺失方式 | 400 |
| 18.3.1 分对数分析过程的参数设置 | 354 | 20.1.2 缺失值处理方法 | 400 |
| 18.3.2 实例分析 | 357 | 20.2 SPSS 缺失值分析 | 404 |
| 18.4 选择模型 (Model Selection) 过程 | 360 | 20.2.1 缺失值分析过程的参数设置 | 404 |
| 18.4.1 选择模型分析过程的参数设置 | 360 | 20.2.2 实例分析 | 408 |
| 18.4.2 实例分析 | 362 | 第 21 章 决策树模型 | 414 |
| 第 19 章 时间序列分析 | 365 | 21.1 决策树模型概述 | 414 |
| 19.1 时间序列概述 | 365 | 21.1.1 CHAID 算法 | 416 |
| 19.1.1 时间序列的组成部分 | 365 | 21.1.2 Exhaustive CHAID 算法 | 417 |
| 19.1.2 时间序列的数学模型 | 366 | 21.1.3 CRT 算法 | 417 |
| 19.1.3 时间序列的分析步骤 | 368 | 21.1.4 QUEST 算法 | 418 |
| 19.1.4 SPSS 时间序列分析功能 | 368 | 21.2 决策树的参数设置 | 418 |
| 19.2 时间序列数据的预处理 | 375 | 21.2.1 变量设置 | 418 |
| 19.2.1 缺失值替换 | 375 | 21.2.2 类别 (Categories) 设置 | 419 |
| 19.2.2 定义时间变量 | 376 | 21.2.3 输出 (Output) 设置 | 420 |
| 19.2.3 时间序列预测的平稳化 | 376 | 21.2.4 验证 (Validation) 设置 | 422 |
| 19.3 指数平滑模型过程 | 377 | 21.2.5 保存 (Save) 设置 | 423 |
| 19.3.1 指数平滑的基本原理 | 377 | 21.2.6 条件 (Criteria) 设置 | 424 |
| 19.3.2 指数平滑模型分析过程的参数设置 | 380 | 21.2.7 CHAID 算法设置 | 425 |
| 19.3.3 实例分析 | 381 | 21.2.8 CRT 算法设置 | 425 |
| 19.4 ARIMA 模型 | 386 | 21.2.9 QUEST 算法设置 | 426 |
| | | 21.2.10 修剪 (Pruning) 设置 | 426 |

| | | | |
|---|-----|-------------------------------|-----|
| 21.2.11 替代变量 (Surrogates) 设置 | 427 | 22.3 实例分析 | 456 |
| 21.2.12 选项 (Options) 设置 | 427 | 22.3.1 参数设置 | 457 |
| 21.2.13 错误分类成本设置 | 428 | 22.3.2 结果分析 | 459 |
| 21.2.14 利润 (Profits) 设置 | 428 | 第 23 章 信用风险分析 | 464 |
| 21.2.15 先验概率 (Prior Probabilities) 设置 | 429 | 23.1 信用风险概述 | 464 |
| 21.2.16 实例分析 | 430 | 23.1.1 信用风险基本概念 | 464 |
| 21.2.17 模型建立 | 430 | 23.1.2 信用风险度量方法 | 465 |
| 21.2.18 模型评估 | 432 | 23.1.3 SPSS 中信用风险分析模块 | 468 |
| 第 22 章 神经网络 | 439 | 23.2 实例分析 | 468 |
| 22.1 神经网络概述 | 439 | 23.2.1 二元 Logistic 分析过程 | 468 |
| 22.1.1 历史及现状 | 440 | 23.2.2 决策树分析过程 | 474 |
| 22.1.2 神经网络特点 | 441 | 23.2.3 判别式分析过程 | 479 |
| 22.1.3 神经元模型 | 442 | 第 24 章 SPSS 在社会经济综合评价中 | 484 |
| 22.1.4 神经网络模型 | 443 | 24.1 沿海省市经济综合指标的主成分 | 484 |
| 22.1.5 神经网络的学习规则 | 443 | 24.2 中国内地城镇居民消费结构的聚类 | 488 |
| 22.1.6 SPSS 神经网络模型 | 444 | 24.3 我国内地可支配收入和消费性支出 | 492 |
| 22.2 SPSS 神经网络模型的设置 | 447 | | |
| 22.2.1 多层感知器 (MLP) 分析 | 447 | | |
| 22.2.2 径向基函数 (RBF) 分析过程 | 454 | | |



第 1 章 SPSS 软件概述

回顾 SPSS 软件的发展历程，从最初的“社会科学统计软件包”(Solutions Statistical Package for the Social Sciences)到 2000 年的 SPSS (Statistical Product and Service Solutions, 统计产品与服务解决方案)软件，SPSS 软件都发生着巨大的变化。IBM 公司于 2009 年 7 月 28 日宣布将用 12 亿美元收购分析软件提供商 SPSS 公司，如今 SPSS 已发布 SPSS 24.0 版本，也标志着 SPSS 的战略方向正在做出重大调整。本章将讲述 SPSS 的发展历程，并介绍 SPSS 基本的使用方法。



本讲内容

- SPSS 24.0 简介
- SPSS 24.0 软件安装、启动及退出
- SPSS 24.0 软件基本操作环境
- SPSS 24.0 帮助系统

1.1 SPSS 简介

SPSS 是英文名称的首字母缩写，英文全称为 Statistical Package for the Social Sciences，即“社会科学统计软件包”。但是随着 SPSS 产品服务领域的扩大和服务深度的增加，SPSS 公司已于 2000 年正式将英文全称更改为 Statistical Product and Service Solutions，即“统计产品和服务解决方案”，标志着 SPSS 的战略方向正在做出重大调整。

SPSS 是世界上最早的统计分析软件，由美国斯坦福大学三位研究生于 20 世纪 60 年代末研制，同时成立了 SPSS 公司，并于 1975 年在芝加哥组建了 SPSS 总部。1984 年，SPSS 总部首先推出了世界上第一个统计分析软件微机版本 SPSS/PC+，开创了 SPSS 微机系列产品的开发方向，极大地扩充了它的应用范围，并使其能很快地应用于自然科学、技术科学和社会科学的各个领域，世界上许多有影响的报刊杂志纷纷就 SPSS 的自动统计绘图、数据的深入分析、使用方便、功能齐全等方面给予了高度的评价与称赞。迄今 SPSS 软件已有 30 余年的成长历史。全球约有 25 万家产品用户，它们分布于通信、医疗、银行、证券、保险、制造、商业、市场研究、科研教育等多个领域和行业，是世界上应用最广泛的专业

统计软件。在国际学术界有条不成文的规定,即在国际学术交流中,凡是用 SPSS 软件完成的计算和统计分析,可以不必说明算法,由此可见其影响之大和信誉之高。

1994 年至 1998 年,SPSS 公司陆续购并了 SYSTAT 公司、BMDP 软件公司、Quantime 公司、ISL 公司等,并将各公司的主打产品收纳 SPSS 旗下,从而使 SPSS 公司由原来的单一统计产品开发与销售转向企业、教育科研及政府机构提供全面信息统计决策支持服务,成为了走在最新流行的“数据仓库”和“数据挖掘”领域前沿的一家综合统计软件公司。

SPSS 是世界上最早采用图形菜单驱动界面的统计软件,它最突出的特点就是操作界面极为友好,输出结果美观漂亮。它将几乎所有的功能都以统一、规范的界面展现出来,使用 Windows 的窗口方式展示各种管理和分析数据方法的功能,对话框展示出各种功能选择项。用户只要掌握一定的 Windows 操作技能,粗通统计分析原理,就可以使用该软件为特定的科研工作服务。SPSS 是非专业统计人员的首选统计软件,在众多用户对国际常用统计软件 SAS、BMDP、GLIM、GENSTAT、EPILOG、MINITAB 的总体印象分的统计中,其诸项功能均获得最高分。SPSS 采用类似 Excel 表格的方式输入与管理数据,数据接口较为通用,能方便地从其他数据库中读入数据。其统计过程包括常用的、较为成熟的统计过程,完全可以满足非统计专业人士的工作需要。输出结果十分美观,存储时则是专用的 SPO 格式,可以转存为 HTML 格式和文本格式。对于熟悉老版本编程运行方式的用户,SPSS 还特别设计了语法生成窗口,用户只需在菜单中选好各个选项,然后单击“粘贴”按钮就可以自动生成标准的 SPSS 程序,极大地方便了中、高级用户。

由上面的叙述可知,SPSS 具有以下特点。

- 操作简便:以对话框方式操作,绝大多数操作过程可通过单击鼠标完成。
- 在线帮助方便:用户可在 SPSS 的任一过程中获得帮助,查询主题和索引,根据帮助框中的指导进行操作。
- 数据转换功能较强:可存取和转换多种数据类型,如 dBase, Lotus, Excel, ASCII 文件等。
- 数据管理功能强大:集数据输入、转换、检索、管理、统计分析、作图、制表及编辑功能于一身。
- 程序生成简化:系统能将对话框指定的命令、子命令和选择项等内容自动编写成 SPSS 命令语句,并可以编辑,继而形成 SPSS 环境下的可执行程序文件。
- 统计分析方法全面丰富:含有最新的统计方法,如对应分析(Correspondence Analysis),联合分析(Conjoint Analysis),多分类变量的 Logistic 回归分析等,且所用方法具有权威性。
- 结果输出规范:输出结果主要为图形方式,规范而简洁,还可根据个人要求编辑输出方式。

1.2 SPSS 操作入门

本节将详细讲述 SPSS24.0 软件的安装、系统环境和基本操作问题,初步展示 SPSS24.0 系统的各种特点和用法。

1.2.1 软件安装、启动及退出

1. SPSS24.0 的安装

首先,从 SPSS 的官方网站上下载 SPSS24.0 软件的安装程序,然后解压到 E 盘中,双击“setup.exe”安装文件;或者把装有 SPSS24.0 软件的安装盘,放入计算机的光驱中,系统会自动的弹出 SPSS24.0 的安装对话框,然后根据对话框的提示即可完成 SPSS24.0 软件的安装。

2. SPSS24.0 的启动

SPSS24.0 软件的启动有两种方式,第一种是双击桌面的快捷方式;第二种是在开始菜单中选择“程序 IBM SPSS Statistics IBM SPSS Statistics 24”,单击即可,如图 1-1 所示。

双击桌面 SPSS24.0 的快捷方式,系统进入如图 1-2 所示的启动界面,则 SPSS 正常启动。进入 SPSS24.0 界面后系统弹出如图 1-3 所示的 SPSS24.0 的启动选项,图 1-3 中各个选项的含义如下。

- 运行图形化教程 (Run the Tutorial)
- 输入数据 (Type in Data)
- 运行存在的查询文件 (Run an Existing Query)
- 新建数据库查询 (Create New Query using Database Wizard)



图 1-1 开始菜单中 SPSS 的选项

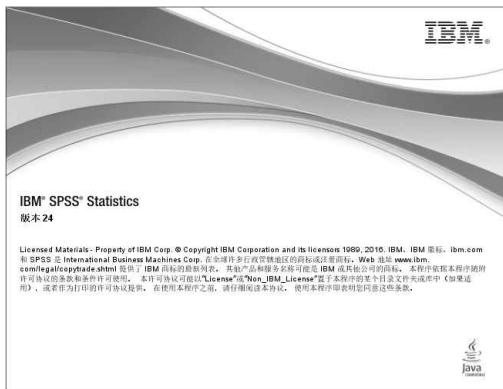


图 1-2 SPSS24.0 启动界面



图 1-3 SPSS24.0 的启动选项

- 打开数据文件 (Open an Existing Data Source)
- 打开类型的数据文件 (Open another Type of File)
- 不要显示此对话框 (Don't Show this Dialog in the Future)

单击图 1-3 中的“确定 (OK)”按钮,则系统进入 SPSS24.0 的主界面,如图 1-4 所示。

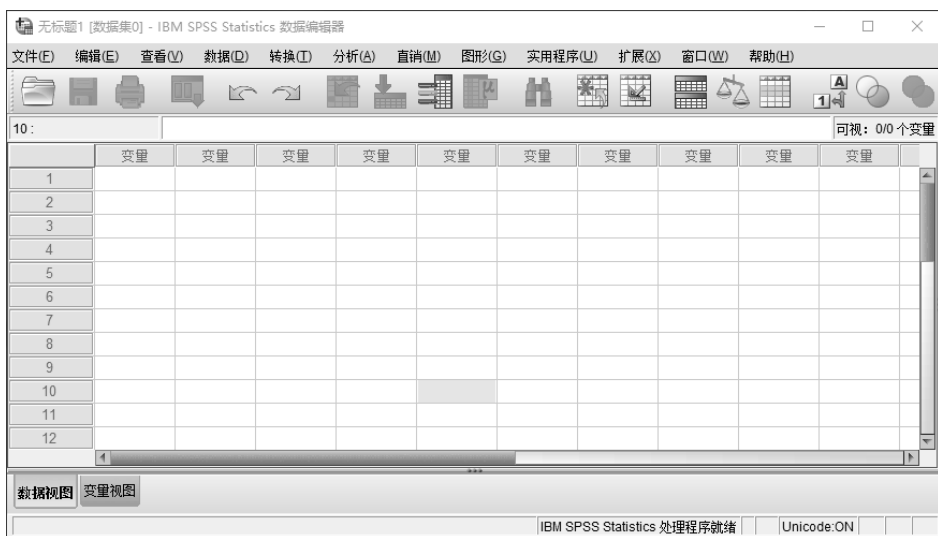


图 1-4 SPSS24.0 的主界面

3. SPSS24.0 的退出

单击“文件”菜单中的“退出”选项，或者直接单击图 1-4 窗口右上角的“关闭”按钮，即可退出 SPSS24.0 软件系统。

1.2.2 操作环境

熟悉 SPSS 软件的操作环境是进行数据分析挖掘的基础，SPSS 的操作环境主要有数据编辑窗口（SPSS Data Editor）、结果浏览窗口（SPSS Viewer）、程序编辑窗口（SPSS Syntax Editor）和脚本编辑窗口（Script）。

1. 数据编辑窗口

启动 SPSS 以后进入的界面，如图 1-4 所示，即为 SPSS 软件系统的数据编辑窗口，此窗口是 SPSS 中最基本的界面，在此窗口中进行数据挖掘前的数据整理编辑工作，也是数据挖掘中非常重要的一步，在图 1-4 中显示的是“数据视图（Data View）”，由此图的底部可以看出，如单击“变量视图”标签，则界面会转换到“变量视图（Variable View）”，如图 1-5 所示。

数据视图窗口：按照行列形式在窗口中显示数据，可以在此窗口中浏览、修改数据值和数据值标签。

- 行：表示观察个体，由观察对象的所有属性组成。
- 列：表示变量，一个变量是所有观察对象的某个属性的集合。
- 数据格：表示对应观察对象的某个属性的观察值或者标签。

变量视图窗口：创建、显示和修改数据视图窗口中变量属性的窗口。

- 行：表示变量。
- 列：表示变量的属性。

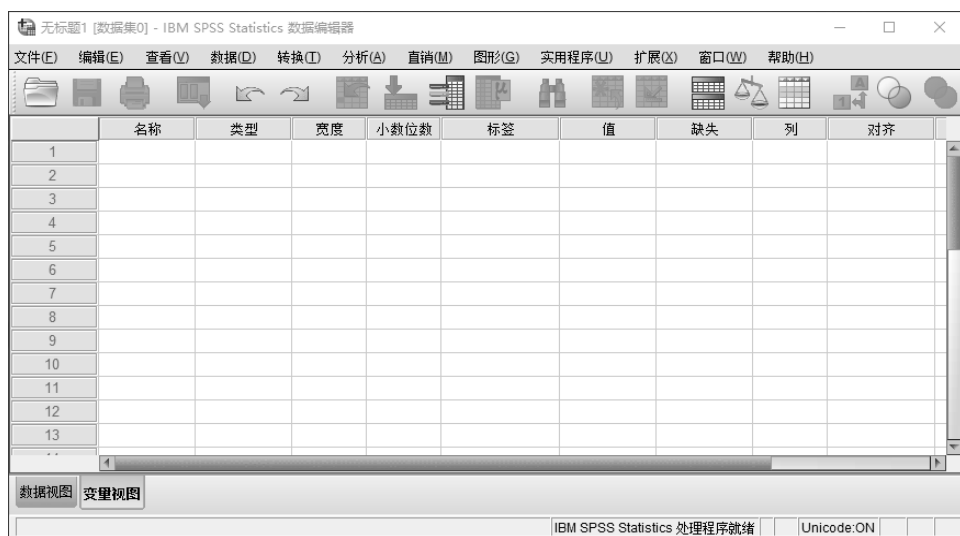


图 1-5 数据变量视图

2. 结果浏览窗口

输出窗口的显示是在选择了一定的变量和选择统计方法或输入了 SPSS 程序命令后才自动生成的。如果运行程序发生错误，则系统给出出错的信息并停止运行。输出窗口主要显示统计结果，包括各种图表等，如图 1-6 所示。



图 1-6 SPSS 的输出窗口

图中的具体内容如下。

- 输出窗口包括两部分：左边为大纲视图，右边为显示统计结果。
- 此结果可以作为输出文件进行保存。
- 输出窗口有自己的菜单栏，其大部分菜单与主菜单相同，输出窗口的菜单也可以执行所有的统计分析功能，对数据文件进行分析，分析结果直接显示在输出窗口。

- 程序中打开多个输出窗口，新开的输出窗口按先后顺序分别标记为输出 1 (output1)，输出 2 (output2) 等。
- 双击输出窗口的生成图形可以进一步对其进行编辑或修改。

3. 程序编辑窗口

单击图 1-4 中的菜单“文件 (File) 新建 (New) 语法 (Syntax)”，即可打开“语法编辑 (Syntax)”窗口，如图 1-7 所示。语法编辑窗口就是编写、调试和运行 SPSS 程序的出口，大部分的 SPSS 功能可以利用窗口操作来完成，通过 SPSS 程序，用户可以获得想要的数据分析过程。



图 1-7 “语法编辑”窗口

语法编辑窗口详细介绍如下。

- 语法编辑窗口按照 SPSS 规则编写 SPSS 程序语句，是一个非激活窗口。只有调开了一个具体的统计分析程序，并通过单击“粘贴 (Paste)”按钮后，此窗口才会打开。
- 在窗口中可以对其内容进行修改、保存，从主菜单中单击“运行 (Run)”按钮可以提交系统运行。
- 其中大部分菜单与主菜单相同，且窗口的菜单也可以执行所有的统计分析功能，对数据文件进行分析，分析结果直接显示在输出窗口。
- 程序中打开多个语句窗口，新开的语句窗口按先后顺序分别标记为语法 1 (Syntax1)，语法 2 (Syntax2) 等。

4. 脚本编辑窗口 (Script)

单击图 1-7 中的菜单“文件 (File) 新建 (New) 脚本 (Script)”，即可打开“脚本编辑窗口 (Script)”，如图 1-8 所示。脚本编辑窗口是一个非常有特色的窗口，其使用 Sax BASIC 语言的编程环境。脚本编辑窗口功能如下。

- 定制输出特征：显示、操作对话框；使用命令语句执行数据转换和统计分析；将图表输出为多种图表格式文件等。

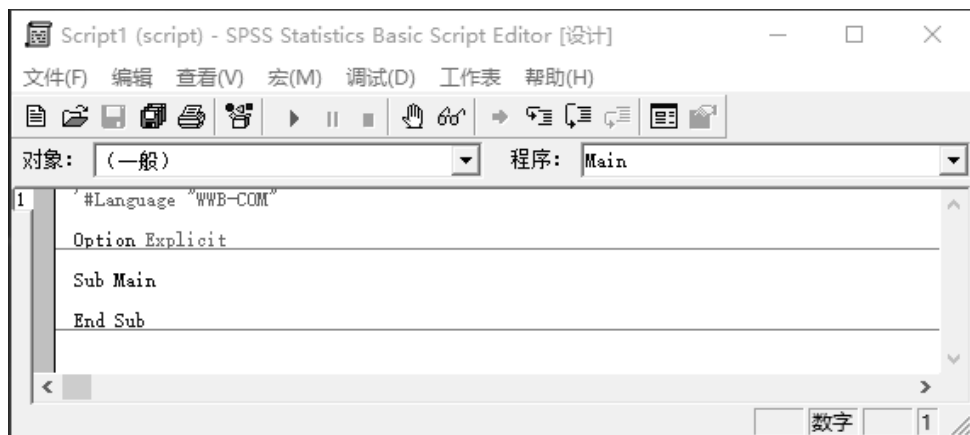


图 1-8 “脚本编辑”窗口

- 通过 Sax BASIC 语言，用户编写自己所需要的程序。
- 在脚本文件夹中安装有较多示范性脚本文件，可以直接调用这些文件来实现某些功能，也可以用这些现存的脚本文件为基础，通过编辑，以实现某些其他功能。
- 程序中可以打开多个窗口，新开的脚本编辑窗口按先后顺序分别标记为脚本 1 (Script1)，脚本 2 (Script2) 等。

1.2.3 系统参数的设置

系统参数设置通过选择主菜单的“编辑”“选项”，然后打开对话框完成，如图 1-9 所示为“系统参数设置”对话框，有些参数设置在设置完成后立即生效，有些要在 SPSS 重启后生效。系统参数设置包括常规、语言、查看器、数据、货币、输出、图表、透视表、文件位置、脚本、多重插补、语法编辑器的参数设置。

1. 常规

常规面板可以设置系统中各种通用参数，所有设置的参数可以自动保存，如图 1-9 所示。各项参数的具体含义如下。

- 变量列表：设置显示变量顺序的方式。下面的单项选择可以设定变量在变量表中的显示方式和显示顺序。显示方式可选变量标签或变量名。显示顺序可选按变量的字母顺序排列或按在文件中出现的先后顺序排序。
- 角色：在支持基于定义的角色预先选择分析变量功能的对话框中，可以进行预定义角色和自定义分配之间的切换。
- 最大线程数：可以选择自动设置数值或者自定义数值。
- 输出选项：包括系统度量 and 结果通知方式等。
- 窗口 (Windows)：启动 SPSS 时语句窗口状态。



图 1-9 “系统参数设置”对话框

2. 语言

语言面板可以设置 SPSS 软件中的界面语言、字符编码和语言环境等，如图 1-10 所示。各项参数的具体含义如下。

- 语言：可以设定输出语言和用户界面语言。
- 数据和语法的字符编码：确定读写数据文件和语法文件的编码方式的缺省行为。
- 双向文本：控制整段文本的文本排列。



图 1-10 “语言选项设置”对话框

3. 查看器

单击图 1-9 中上方的“查看器”选项，则打开如图 1-11 所示的“查看器”对话框。在改变了参数以后，再次运行 SPSS 后才能生效，如图 1-11 所示。

各项参数的具体含义如下。

- 设置输出状态：控制每次运行中自动显示或隐藏的项，以及各项的初始对齐方式。
- 标题：用于输出结果标题的文字设置。
- 页面标题：实现对文本输出的界面设置。
- 文本输出：文本输出设置。
- 缺省页面设置：控制用于打印的方向和页边距缺省选项。



图 1-11 “查看器”对话框

4. 数据

单击图 1-9 中上方的“数据”选项，则打开如图 1-12 所示的“数据窗口设置”对话框。主要设置一些数据处理过程的更新方式、新变量显示格式、日期格式和随机数生成等参数。

- 数据的转换和合并选项。
- 显示新的数值变量的格式。
- 随机数生成器的设置。
- 设置用两位数字表示年号的年限全距。
- 指定测量级别。



图 1-12 “数据窗口设置”对话框

5. 货币

单击图 1-9 中上方的“货币”选项，打开后的对话框如图 1-13 所示。

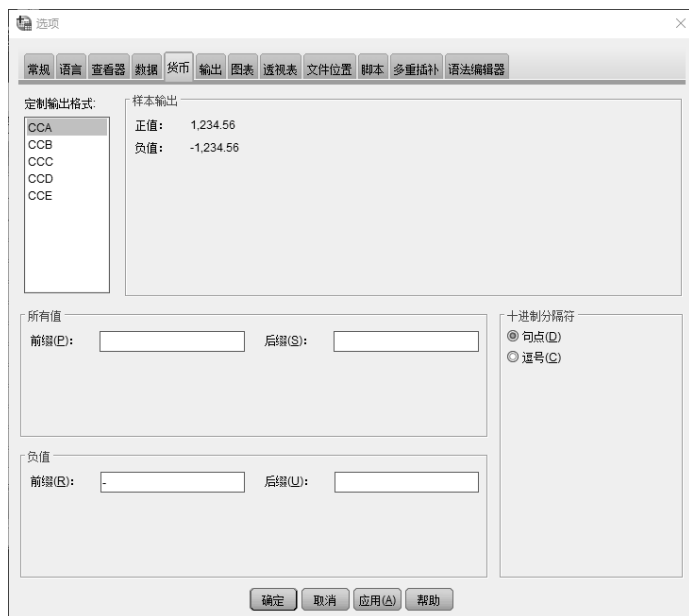


图 1-13 “货币选项设置”对话框

各项参数的具体含义如下。

- 定制输出格式：用户定义输出栏，可以设置 5 种自定义的格式，分别命名为 CCA、CCB、CCC、CCD 和 CCE。

- 样本输出。
- 所有值：设置数值的首尾字符。 前缀（Prefix），加入前缀字符；后缀（Suffix），加入后缀字符。
- 负值：设置负数的首尾字符。
- 十进制分隔符。

6. 输出

单击图 1-9 中上方的“输出”选项，则打开如图 1-14 所示的“输出选项设置”对话框，输出选项控制某些输出选项的缺省设置。

- 大纲标注：设定输出时是否使用标签。包括设置窗格中的变量名称、变量标签、数据值和值标签显示。
- 透视表标签（Pivot Table Labeling）：在要点表格中，设定输出表格时是否使用标签。
- 单击描述。
- 输出显示。
- 屏幕阅读器辅助功能：设置屏幕阅读器如何朗读透视表行标签和列标签。



图 1-14 “输出选项设置”对话框

7. 图表

单击图 1-9 中上方的“图表”选项，则打开如图 1-15 所示的“图表输出设置”对话框。

- 图表模板（Chart Template）：可以使用当前设置的各种参数，也可以使用保存在模板文件中的参数建立新输出的图形。
- 当前设置（Current Settings）：文本风格设置。

- 框架 (Frame): 图形边框设置。
- 网格线 (Grid Lines): 图形网格线。
- 样式循环 (Style Cycles): 包括颜色、线段等。



图 1-15 “图表输出设置”对话框

8. 透视表

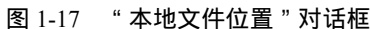
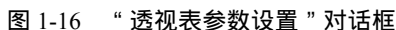
单击图 1-9 中上方的“透视表”选项，则打开如图 1-16 所示的“透视表参数设置”对话框。

- 表格外观 (Table Look): 选择应用表格的外观样式，选中样式会显示在右边的样本 (Sample) 栏中。单击“应用”或“确定”按钮。浏览 (Browse) 表示从其他目录中选表格外观文件，直接设置表格外观 (Set Table Look Directory) 表示选择系统默认的表格外观目录。
- 列宽度 (Column Width): 控制表格列宽。
- 表格注释。

9. 文件位置

单击图 1-9 中上方的“文件位置”选项，则打开如图 1-17 所示的“本地文件位置”对话框。

- 打开和保存对话框的启动文件夹 (Startup Folders for Open and Save dialogs)
- 会话日志 (Session Journal): 所有运行的命令将保存在一个日志文件里，包括附加模式 (Append) 和覆盖模式 (Overwrite) 两种保存方式。



单击图 1-9 中上方的“脚本”选项，则打开如图 1-18 所示的“脚本编辑窗口设置”对话框。自动脚本栏（Autoscript），即脚本子程序的组合。具体包括基础自动脚本（Base Autoscript）和单一对象自动脚本（Autoscript for Individual Objects）。

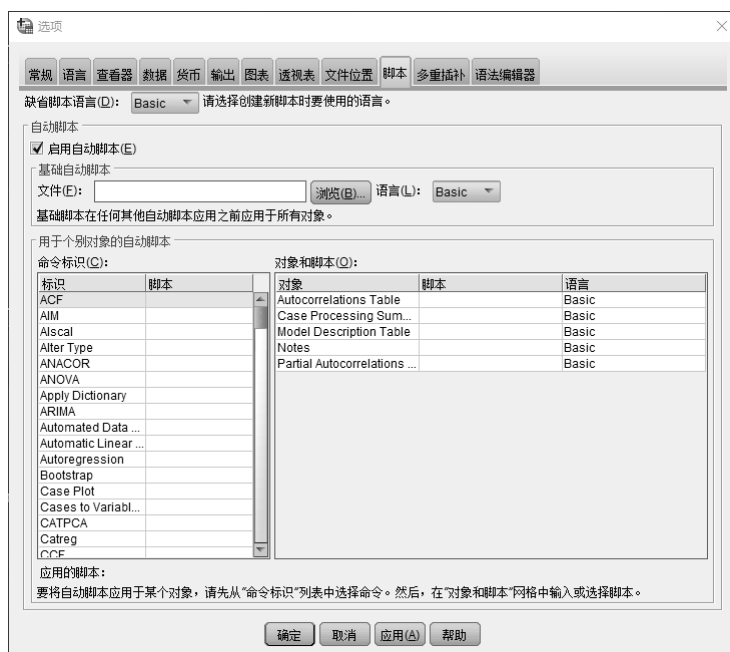


图 1-18 “脚本编辑窗口设置”对话框

11. 多重插补

单击图 1-9 中上方的“多重插补”选项，则打开如图 1-19 所示的“多重插补窗口设置”对话框。其中包括归因数据标记和分析输出设置选项。



图 1-19 “多重插补窗口设置”对话框

12. 语法编辑器

单击图 1-9 中上方的“语法编辑器”选项，则打开如图 1-20 所示的“语法编辑器窗口设置”对话框。其中包括显示语法颜色编码、错误颜色编码、自动完成设置、装订线、窗格等选项设置。



图 1-20 “语法编辑器窗口设置”对话框

1.3 SPSS 的帮助系统

在学习 SPSS 软件的操作过程中，一定会遇到各种各样的问题，因此，快速地解决问题对提高学习 SPSS 的效率至关重要，在 SPSS 系统中提供了功能非常全面的帮助系统，在 SPSS24.0 主界面的菜单中，Help 帮助系统包含了大部分的帮助信息，包括帮助电子书和 SPSS 各种案例教程。如图 1-21 所示为“帮助”菜单内容，其中各选项如下。

- 主题 (Topic) 选项。
- 教程 (Tutorial) 选项。
- 个案研究 (Case Studies)。
- 统计辅导 (Statistic Coach)。
- 指令语法参考 (Command Syntax Reference)。
- SPSS 社区 (SPSS Community)。

除了上述帮助系统外，单击“帮助”菜单中的“IBM SPSS 产品主页”选项，则会进入 SPSS 官方网站，在其网站上有很多学习资料，用户可以学习其官方网站上的学习资料。

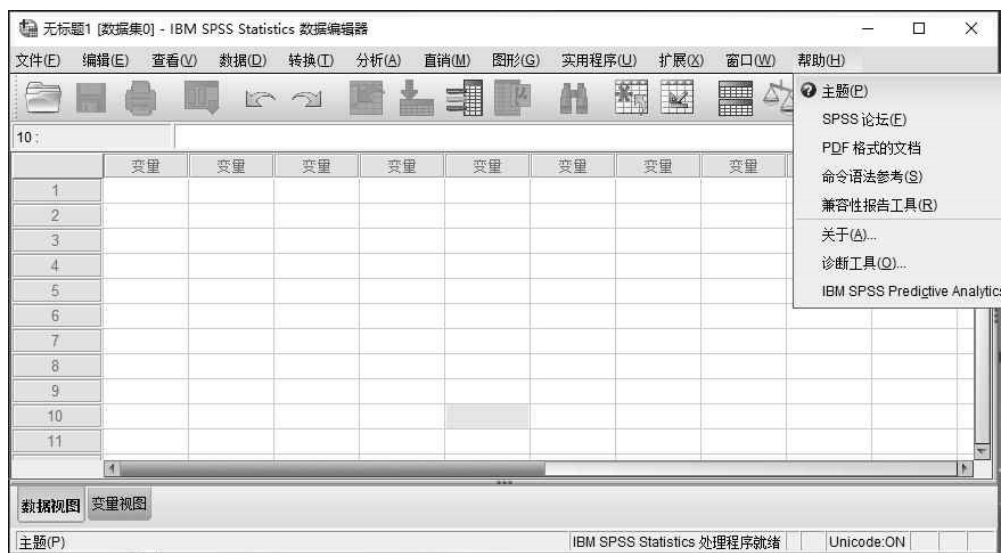
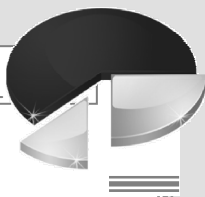


图 1-21 Help 帮助系统

当进行数据挖掘的分析操作时候,例如,用户在进行聚类分析的时候,首先要进入聚类分析的主界面,如图 1-22 所示。当用户在进行参数设置的时候,可以单击图 1-22 中右下方的“帮助 (Help)”按钮,直接进入帮助系统。



图 1-22 聚类分析的主界面



第2章 SPSS 数据挖掘系统

每个企业每天都会产生大量的数据，这些数据来自不同的数据源。数据挖掘是对大量的原始数据进行选择、分析和建模，从中发现以前没有发现的趋势和模式。数据挖掘采用与文本文件同样的分析技术。通过数据和文本挖掘得到的信息对企业战略决策有很大帮助。因此，数据挖掘使得人们从原始数据发展到更加明智的业务决策。本章将详细叙述数据挖掘的过程、工具、用法，以及 SPSS 软件系统在数据挖掘中的位置等，为利用 SPSS 软件系统进行数据挖掘打下基础。



本讲内容

- 数据挖掘概述
- 如何进行成功的数据挖掘
- SPSS 数据挖掘模块介绍

2.1 数据挖掘概述

你是否正在做数据挖掘项目？你是否正在考虑要第一次承担一个数据挖掘项目？无论你是否具有数据挖掘经验，SPSS 数据挖掘系统，都可以为你在计划和实施数据挖掘项目时，提供方便的指导。

在数据挖掘过程中应用 SPSS 数据挖掘软件，用它来节省金钱，及时实施项目并得到最好的结果。

2.1.1 数据挖掘的含义

数据挖掘是按照既定的业务目标，对大量的企业数据进行探索，揭示隐藏在其中的规律性并进一步模型化的先进、有效的方法。数据挖掘是在大型数据中发现隐含模式和关系的过程，数据挖掘解决了普通的疑惑。例如，你拥有越多的顾客信息，有效地分析和得出有意义的结论则越困难，越耗时间。数据有大量有价值的信息，却经常由于缺少才智、时间或技术，而未被开发。数据挖掘是用一个清晰的商务定向和强大的分析技术来快速、完全地挖掘山一样的数据，取出有价值的、有用的信息——你所需要的商务才智。

当然,并非所有的信息发现任务都被视为数据挖掘。例如,使用数据库管理系统查找个别的记录,或通过因特网的搜索引擎查找特定的 Web 页面,则是信息检索(Information Retrieval)领域的任务。虽然这些任务是重要的,可能涉及使用复杂的算法和数据结构,但是它们主要依赖传统的计算机科学技术和数据的明显特征来创建索引结构,从而有效地组织和检索信息。尽管如此,数据挖掘技术也已用来增强信息检索系统的能力。

2.1.2 数据挖掘与 OLAP

在比较成熟的系统中,数据分析过程都是基于以数据仓库为基础,OLAP(On-Line Analytical Processing,在线分析处理)和数据挖掘相辅相成的分析模式。数据仓库将来自于各种数据源的数据,根据不同的主题进行存储,并对原始数据进行抽取、转换和加载等一系列筛选和清理工作。OLAP 则将数据通过多维视角和多种层次向用户进行多方式的呈现。数据挖掘则应用不同的算法,向用户揭示数据间的规律性,从而辅助商业决策。

报告和 OLAP 是用于理解过去所发生事务的重要工具。数据挖掘则是用于了解将来可能发生事务的方法。数据挖掘用预测性模型(包括统计和机器学习技术——如神经网络)来预测将来。

例如,查询和报告告诉你上个月的总体销售情况。OLAP 则层层深入地告诉你上个月各项产品的销售情况。然而,数据挖掘会告诉你下个月谁可能会买你的产品。而且,为了最好的商务效益,将开发和数据挖掘相结合,以发现如何使产品个性化以导致最大可能性的购买。

OLAP 和数据挖掘的主要区别在于:在辅助决策时,前者是基于用户建立的一系列假设驱动,通过 OLAP 来证实或者推翻这些假设,是一个演绎推理的过程;数据挖掘是通过归纳的方式,在海量数据中主动找寻模型,自动发掘隐藏在数据中的价值信息。

相对于 OLAP,数据挖掘把更多的主动权交给了挖掘工具,在一定程度上,可以看成人工智能的初级应用。此外,OLAP 限于结构化数据,侧重与用户的交互、快速响应,以及提供多维视图,而数据挖掘还可以分析诸如文本的、空间的和多媒体的非结构化数据。

二者相辅相成。OLAP 的分析结果可以补充到系统知识库中,给数据挖掘提供分析信息并作为数据挖掘的依据;数据挖掘所发现的知识可以指导 OLAP 的分析处理,拓展 OLAP 的深度,以便发现 OLAP 所不能发现的、更为复杂而细密的信息。

2.1.3 数据挖掘和统计学

数据挖掘并不是对统计学的代替。实际上,统计学是对数据挖掘的很好的补充。经典的统计学技术,如回归与数据挖掘技术、神经网络一起应用。统计学也可用于验证数据挖掘结论。

显然,统计学和数据挖掘有着共同的目标:发现数据中的结构。事实上,由于它们的目标相似,一些人(尤其是统计学家)认为数据挖掘是统计学的分支。这是一个不切合实际的想法。因为数据挖掘还应用了其他领域的思想、工具和方法,尤其是计算机学科,如数据库技术和机器学习,而且它所关注的某些领域和统计学家所关注的有很大不同。

相对于统计学而言，准则在数据挖掘中起着更为核心的作用并不奇怪，数据挖掘所继承的学科，如计算机科学及相关学科也是如此。数据集的规模常常意味着传统的统计学准则不适合数据挖掘问题，不得不重新设计。当数据点被逐一应用以更新估计量，适应性和连续性的准则常常是必需的。尽管一些统计学的准则已经得到发展，但更多的应用是机器学习（正如“学习”所示的那样）。

另外，统计学很少会关注实时分析，然而数据挖掘问题常常需要这些。例如，银行事务每天都会发生，没有人能等三个月得到一个可能的欺诈的分析。类似的问题发生在总体随时间变化的情形。研究组有明确的例子显示银行债务的申请随时间、竞争环境、经济波动而变化。

2.1.4 数据挖掘的目的

当有一个对商务目标可靠、详细的指导时，那在一些商务决策上就有能力做出正确的决定。数据挖掘，通过了解过去和现在，得出准确的预测，使人有能力掌握和改变公司的命运。例如，数据挖掘会告诉哪些人可能会变成有利可图的消费者，哪些人是最可能带来回报的。有了对未来的了解，可以通过只为那些可能带来回报的、有价值的消费者提供服务来增加 ROI。这些决定是基于合理的商务才智，而不是基于本能的反应。这些决定会带来一致的结果，使其处于竞争的前列。

2.1.5 数据挖掘应用

可以用数据挖掘来解决几乎所有与数据相关的商务问题，一般较常见的应用案例多发生在零售业、直效行销界、制造业、财务金融保险、通信业，以及医疗服务等，例如：

- 从消费者处增加收入。
- 了解消费者的划分和优先顺序。
- 识别有利可图的消费者，争取新的顾客。
- 交叉销售和提升销售。
- 保持消费者和提高忠诚度。
- 增加 ROI 和降低改进成本。
- 检测欺骗、浪费和滥用。
- 检测信贷风险。
- 增加网络地址收益性。
- 增加商店交易、最优化摆放产品以增加产品销售。
- 追踪商业成绩。

2.1.6 数据挖掘流程

SPSS 数据挖掘产品和服务通过支持交叉行业数据挖掘标准流程（Cross-Industry Standard Process for Data Mining, CRISP-DM）而保证及时、可靠的结果。由工业专家创造

的 CRISP-DM 为数据挖掘过程的每一阶段的任务和目标提供指导。CRISP-DM 是工业标准化的数据挖掘过程。

CRISP-DM 阶段包括以下几个部分。

- 商业理解：明确了解所面临的商务挑战。
- 数据理解：决定什么数据可以用于数据挖掘，以得到答案。
- 数据准备：以合适的格式来准备数据，回答商务问题。
- 建立模型：设计数据模型来满足要求。
- 模型评估：用结果逆向检测项目目标。
- 成果发布：使项目结果有助于决策者。

2.2 成功的数据挖掘

在数据挖掘项目中应用 CRISP-DM 指导方针会起很大作用，它会引导一个成功的数据挖掘项目。跟随一个已证实的方法是关键的——在商务问题中，复杂的数据挖掘技术和大量的可用数据会把没有稳固基础的项目压倒。

例如，没有专门的处理，可能会掉进通常所说的陷阱：“让我们通过某些数据挖掘运算法则来处理数据，看看我们可以得到什么结果。”结果是数据模式可能并不适用于你的情况。

其次，为了能在项目结束时显示 ROI，在开始前要了解将要怎样评估结果。以有限的目标和计划开始，当达到成功时，再移向更复杂的计划。

最后，一个数据挖掘项目是需要集体努力的。数据挖掘要求商务用户了解问题和数据，以及懂得分析。也需要为数据拥有者提供入口。例如，需要一个数据挖掘分析者，一个数据库分析者，一个市场经理。这些人可能会因项目目标不能站在一起，而掉进目标的不同功能领域。因此，寻找方法使他们的角色能合作良好是很重要的。

数据挖掘是指一个完整的过程，该过程从大型数据库中挖掘先前未知的，有效的，可实用的信息，并使用这些信息做出决策或丰富知识。如图 2-1 所示描述了数据挖掘的基本过程和主要步骤。

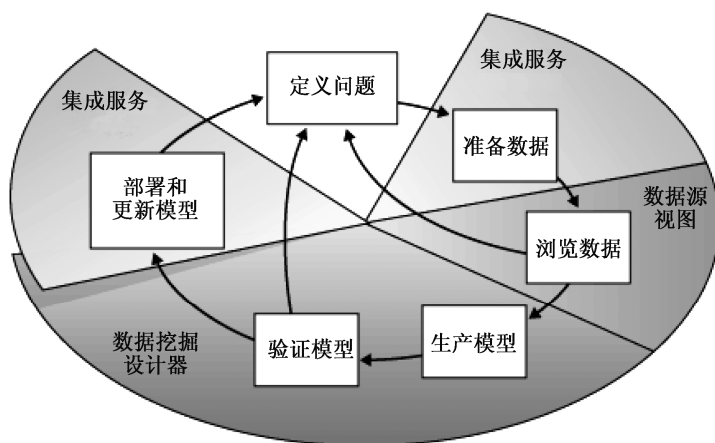


图 2-1 数据挖掘流程图

下面将详细叙述如何进行数据挖掘。

2.2.1 CRISP-DM 方法论

对上面的 CRISP-DM，详细说明各个阶段。

1. 商务理解

从商务角度知道“谁，什么，什么时候，哪个地方，为什么和怎么样”。发展一个对项目参数的完全的理解：目前商务情况，此项目的基本商务目标，成功的标准和谁来决定项目的成功性。

考虑一下如何利用数据挖掘项目的结果。

- 此结果是否将被不需要对结果进行翻译的专家所用。
- 此结果是否将被需要对结果进行不同程度翻译的广大雇员所用。
- 此结果是否将被需要某种格式的某一媒体（如在线，报纸等）所用。

一旦初始的结果完成后，将怎样管理数据呢？如果项目在发展战略过程中，将要做以下工作。

- 周期性的分析新的数据。
- 实时分析新的数据。
- 确保预先了解项目的每一方面以保证拥有成功的必要条件。
- 个体（项目资助者、商务和技术专家）。
- 数据来源。
- 计算机资源。
- 软件。

项目要求。

- 完成计划。
- 结果的易理解性。
- 安全性。
- 数据应用的法律依据。

数据分析的假设。

- 数据的质量（准确性和可利用性）。
- 外部因素（经济论题）。
- 内部因素（商务问题）。
- 模型。

检查和发展下列问题的解决方案。

- 经费、时间、资源等一般性问题。
- 数据资源利用权。
- 数据的技术可利用性。
- 相关知识的可获得性。

确保每位参与者理解整个项目过程中所用的术语和想法。通过制作与该项目相关或

特殊的商务和技术术语词汇表,使部门间更容易地相互理解。

决定必须完成的数据挖掘任务以便达到你的商务目标。用技术术语对数据挖掘任务下定义。例如,商务目标是“增加现有消费者的目录销售量”可能被翻译为数据挖掘目标:“假设有前三年消费者的购买量,相关人口学信息和项目价格,预测消费者会买多少小部件。”

运用技术术语,描述为了考虑项目成功而必须满足的标准。例如,模型显示明确的预测准确性水平,或购买倾向性必须有一个明确的升高度。

制订一个计划,概述你将采取的每一步要达到的数据挖掘目标和要满足的商务目标。评述要用什么工具和技术来使你完成计划。

2. 数据理解

数据理解阶段的首要任务是确保数据是可利用的,否则将一事无成。收集本项目所需要的所有数据。如果你的数据有多个来源,确保你的数据挖掘工具能合并数据。不同来源的数据可能产生质量问题(不同的属性或数值)。关注潜在的问题,约80%的数据在文本文件中可能被掩盖。用文本挖掘工具有效地搜索这些有价值的信息资源。

对数据进行探索性分析:通过分析小量的多来源数据和交流发现,帮助数据库构造者设置优先权。这会有助于你确认重要的区域和潜在的自我实践区域。

数据是否覆盖相关的特征:通过选择最能代表欲分析的环境或行为的数据而确保成功。也决定你的发展选项是否能处理你所有的数据和限制你的数据的实际开发。

描述现有的数据:通过产生一个描述数据格式、记录和区域数目、区域身份和其他相关特征的报告来得到你的数据的清晰图像。

检查数据质量:为了防止出现问题,需要评估数据的质量和对被检测的任何问题做一个计划。

- 属性的名字和它们所包含的值是否一致。
- 是否有属性被遗失。
- 是否有空白区域。
- 检查值的多个拼写以避免重复。
- 寻找偏离和确定原因。
- 检查任何属性的应答是否与常理违背。
- 排除那些不相关的数据。

产生数据质量报告:检查数据副本,潜在的数据错误(例如,在他们变成消费者前,显示他们已经流失)和可能包含错误信息的强制性的数据库区域。

3. 数据准备

首先要选择用于数据分析的数据,选择数据的原因如下。

- 完成显著性和相关性检测来决定包括哪些字段。
- 选择数据子集。
- 利用抽样技术检测小块数据的适当性以决定某些属性是否较其他的数据更为重要,并对其进行加权处理。

描述数据质量:为了确保可靠结果,现在花时间来调适任何数据问题。内容可能包括以下几个方面。

- 怎样处理噪声数据。

- 指出特定值和它们的意义。例如，当一个问题没有被回答或者当数据由于空间的考虑而被截短时，可能会被默认为某一特定值。
- 某些字段可能与目标不相关，也不需要被清除。对这些字段采取跟踪行为或不跟踪行为，因为可能会在后来的过程中决定用它们。

选择一个灵活的数据重构工具：确保所选择的数据挖掘工具能够根据项目需要来构建数据库。你的工具应该允许你增加所需的新的字段。要记住数据挖掘是一项发现驱动型的过程——预先是不可能知道数据会带你走到哪里的。

决定是否产生衍生属性。由于下面的原因，可能要产生衍生属性。

- 由于你对环境的丰富经验，知道尽管某一属性当前不存在，但它是非常重要的。
- 模型运算法则仅仅处理某种数据类型，因此，如果不重新生成变量的话，它不能被包括进去。
- 模型结果揭示：相关事实未被提出。

通过合并数据来巩固信息，当加入新的表来巩固信息时，可能也会想产生新的字段，汇总值。也可能要从非电子数据源（纸张报告、专家意见等）获得数据。

数据挖掘工具是否要求特定的顺序：在此阶段，如果数据挖掘工具要求数据集呈特定的顺序，可能需要整理数据子集。

数据是否应该被平衡：检查模型技术是否需要平衡数据。例如，直邮促销经常返回反应信息偏向“无反应”。某些技术做出无反应预测，因此要有一个高的准确性。然而，为了准确地预测正反应，某些技术可能要求使用大致相等的数字。

4. 模型

为了将数据与正确技术相匹配，检查每种技术对数据格式和质量所作的假设。在某些情况下，只有一种技术适合情况。所以，一定要确保如下几点。

- 什么技术适合分析你的问题。
- 是否有管理期望、易懂性的要求。
- 是否有特定的数据特征、时间等任何限制。

模型构建前的检测：在产生模型前，检测计划应用的技术的质量和正确性。产生一个包含训练测试、测试和确认的检测设计。然后构建训练的模型，用测试数据子集评估它的有效性。

构建模型：为了产生一个模型，将已经准备好的数据子集运行模型工具。描写结果，评估它预期的准确性、有效性和潜在的缺点。产生一个详细的模型报告，报告中列出产生的规则，所用的参数设置，模型技术行为和解释，任何有关数据显示模式的结论。在开发时，在正确场合下，用唯一的可用于模型的属性。例如，如果想产生一个用于预测一年后消费者属性的模型，那么用从一年点到准确及时反映那一点的消费者行为的消费者数据。

模型构建后的检查：确保模型得出有助于达到数据挖掘目标的结果。

试用几种模型以得到正确的拟合：为了改善模型性能，试着增加或删除选项，或用可选项来实验。而且，虽然每项技术可能有轻微的差别，试着变化（如聚类和相关）来找到所有的相关模式。

- 倾向模型对以下是适用的：预测消费者行为——发现谁最可能购买，最可能拖欠贷

款和更多。用这些信息来决定哪些消费者和潜在的消费者提供最长期的利润。

- 聚类模型对以下是适用的：发现有相似特征的案例的自然分组——运用聚类分析将异常的信用卡办理的相似案例分组而检测欺骗行为。
- 关联模型对以下是适用的：市场购物篮分析——揭示哪些产品是最可能被一起购买的。通过分类和货架摆放，推荐火车、电话和直邮服务及其他更多。可提高总的销售率。
- 统计模型对以下是适用的：初始分析——统计分析在数据挖掘项目的早期是有用的，以获得数据结构的总体情况。一个简洁的数据特征的描述有助于项目成员得出假说，计划下一步的分析。

评估数据挖掘结果：检测通过给定模型得出的结论是否有助于达到商务目标。是否因为商务原因而使模型有缺陷。如果时间和资源是可利用的，试着检测模型或在实际应用中检测应用模型。

决定下一步：现在应该决定项目是否足够成功，可以向前开发。如果答案是否定的，做任何必要的步骤，以达到满意的结果。记住以下几点。

- 每一结果的潜在的开发。
- 过程如何得到改善。
- 是否存在一些资源，可以进行额外的步骤，或者重复先前的步骤。

5. 成果发布

取得项目结果，决定如何最好地利用它们来陈述你的商务论题。

- 总结可开发的模型或软件结果。
- 开发和评估替代性开发计划。
- 确保结果怎样被分发给接收者。
- 决定如何监测结果的应用，测量效益。
- 确定开发中可能存在的问题和缺陷。

监测和维持计划：确保数据挖掘结果得到最好的利用。可建立结果维持计划，描述如下。

- 未来什么可能发生改变而影响结果的应用。
- 如何监测结果的准确应用。
- 如果必要，什么时候中断该结果的应用或开发。

依赖于开发计划，报告可能是项目总结或是数据挖掘结果的最终陈述。为了建立最终报告，应做到以下几点。

- 确定需要什么报告（幻灯片，管理总结等）。
- 分析数据挖掘目标如何被满足。
- 确认报告接收人。
- 列出报告的结构和内容。
- 选择所要包括的发现。

6. 实施开发计划

根据开发计划分发数据挖掘结果，通过此计划而对结果选择性利用。如果不被用于改善商务，即使最聪明的发现也不会带来高的投资回报。

2.2.2 选择数据挖掘工具

数据挖掘过程中，数据挖掘的工具同样也是很重要的，下面举几个例子。

寻找一个已被证实的可用于解决项目所陈述商务问题的数据挖掘工具：即选择一种所知道的，可以用于解决公司问题，并在计划应用方面有成功记录的工具。

选择用于在商务理解和数据挖掘技术方面起到沟通作用的工具：确保工具所用的步骤与数据挖掘的商务需要相匹配。

- 工具是否可清晰地表达数据挖掘概念。
- 工具是否与项目管理软件或其他可能用的工具相结合？如果不能，是否不得不新建应用软件以弥补此不足。

确保工具可对现有的数据资源和格式进行操作：如果能选择一种能提取和合并多来源、多格式数据，将会节省时间和金钱，并最大可能地得到可靠结果的工具。这一点很重要，尤其是在数据挖掘过程的后期发现不得不从新的来源加入新的数据时。

寻找交互式开发和可视化能力：选择一种可提供交互式可视技术的工具会使开发和理解数据变得容易。这些技术会使你通过在图内变化及根据不同的数据尺度产生新的图表更快地获得直觉。

选择一种可高效、易解的进行数据准备的工具：选择一种可高效进行数据准备（从初始步骤到模型建立），且以易于理解的方式表达数据准备步骤的工具会节省时间和资源。这会使不同经验水平的项目成员获得有效的结果。

确保工具可自动地提取数据：选择一种可为不同数据步骤自动提取数据的工具，可以避免耗时的人工书写查询。

该工具是否可在合理的时间内建立有效的模型：寻找一种工具，它可使分析家快速找到最有效的模型。这种工具应该支持有效的建立和检测多个模型。

选择一种含宽范围技术的工具：为了确保最好的结果，确保工具能为可视化、分类、聚类、相关和回归提供一个宽范围的技术或运算法则。例如，可能发现，对某一数据而言，一种技术比另一种更好。你需要能灵活地试用多种技术以获得准确、有效的结果。这种工具应该能联合应用在不同情况下可获最佳结果的多种技术。

该工具是否可利用现有的数据和设备：选择一种数据挖掘工具，它能利用现存的数据——或数据库中或文件中，也能与现有的分析和可视化工具相兼容。你不会愿意因为不能利用现有的数据库而浪费时间和资源再新建一个。

选择一种可发送一致的、高质的结果的工具：要得到准确的结果，需利用在各种情形下都能很好工作的、适应性强的数据挖掘环境和各种数据的工具，而不是单一地为某一类型的数据或环境而设计的工具。工具应该能管理任何可能有助于阐释商务问题的数据。

⑪ 什么是工具的开发能力：选择一种能将结果合并入现有的和将来的操作应用中的工具，是很重要的，因此也要考虑如下几点。

- 这种合并是否符合成本收益，或是否需要投入额外的时间和财力。
- 这种工具是否能容易地校正工具挖掘结果，如果可以的话，需要什么样的额外的投资。

⑫ 评估与工具相关的所有权潜在的成本：分析每项工具的潜在投资回报，如实施你的数据挖掘工具需要花多长时间？它是否为技术专家而设计？或者它是否适用于不同专业的

用户？现在和将来的培训成本是什么？这种工具是否专为你的特殊用户和商务需要作了定制？你是否可节省普通的过程和使任务自动化？

2.2.3 SPSS 数据挖掘

SPSS 利用一个完全的数据挖掘方案——从理解商务问题到向决策者发放结果——可使数据挖掘过程呈流线和加速化。SPSS 利用已证实的 CRISP-DM 过程将数据挖掘能力带给各种水平的用户，所以，会得到更多的值和更好地利用结果。利用 SPSS 产品可达到广泛的商务目标。

- 增加商务单元和总利润率。
- 理解顾客期望和需求。
- 确认有利可图的消费者和获得新的消费者。
- 保留消费者及增大忠实度。
- 提高投资回报，降低提升成本。
- 总销售和增长销售。
- 测定信贷风险。
- 提高网址收益性。
- 提高商店交易量，为增加销售而使货架安排最优化。
- 监测商务成绩。

1. SPSS Base

当理解了数据时，需要为分析而对它们进行准备。SPSS Base 是一个服务于分析过程——计划、数据收集，数据获取和管理、分析、报告和开发的，有标准组件、紧密结合、全系列的产品线，也是数据挖掘程序的关键组件。第 1 章中已经详细介绍了 SPSS Base 模块的情况，此处只作简要说明。

首先，SPSS 可以让你更快访问和分析大型数据，并且可以处理其他分析工具无法解决的大型数据，因为 SPSS 事实上完全取消了一般分析工具普遍存在的文件大小限制。无论是使用共用数据库，还是从 Web 下载数据，访问和管理数据都变得比以前更加轻松。

只要把 SPSS 和 SPSS Server（选件）连在一起，就可以让服务器去做繁重的计算工作，从而以尽可能快的速度进行分析。

进行数据分析之前，需要准备数据以便分析。SPSS Base 包含的众多技术和功能特性使数据准备简单易行。利用 SPSS Base，可以轻松地实现数据字典的建立（如值的标签和变量类型），并且利用定义数据属性工具，可使分析前进行的数据准备工作更加快捷。SPSS 使人们能够轻松地识别重复观测，以便在数据分析前删除它们。而且，SPSS 能使分析连续型数据的准备工作简单易行。可以在一个 SPSS 会话同时打开多个数据集，这样既节省时间，又精简了数据文件合并的步骤。这也确保了在多个数据集间，复制数据字典的连贯性等。

在数据分析方面，除了一般常见的摘要统计和行列计算，SPSS Base 还包括在基本分析中最受欢迎的统计功能，如集合、计数、交叉分析、分类、描述性统计、因子分析、回归及聚类分析等，而且还可以把分析结果回写到数据库。

当然，SPSS 在图形用户界面方面也有着巨大的优势，可以很简单地用交互式图表清晰地表达分析结果。利用图表构建程序——SPSS 的全新的图表创建界面，能够更轻松地创建

常用的图表。只要把变量和元素拖到图表创建面板，就可以创建图表。也可以随意地利用库中存在的模板快捷地创建图表，而且也可以同时预览将要生成的图表。利用图形生成语言（GPL），高级用户能够创建更多图表。

在 SPSS OLAP 方面，OLAP 改变了常规的创建和共享信息的方式。与其他 OLAP 系统相比，SPSS Report OLAP 含有更多的分析功能，提供了一个快速、灵活的途径来创建、发布和处理用于特别决策判断的信息。

SPSS 从数据采集、分析到结果的呈现，都做了全新的改进。可以建立个性化的工作界面，通过宏程序来完成反复分析、格式化与报表等工作。只需轻按一下按键，便可自动完成一系列的工作任务。

本书中，将详细介绍 SPSS Base 模块，利用 SPSS Base 模块进行数据挖掘分析。

2. SPSS Clementine

Clementine 是 ISL (Integral Solutions Limited) 公司开发的数据挖掘工具平台。1999 年，SPSS 公司收购了 ISL 公司，对 Clementine 产品进行重新整合和开发，现在 Clementine 已经成为 SPSS 公司的又一亮点。

数据挖掘可以帮助你更清楚地了解企业的现状，更深入地洞察企业的未来。企业使用 SPSS 公司的 Clementine 数据挖掘平台，可以实施数据挖掘项目：通过分析整合的多种类型的数据，获得企业运营全面深入的知识，包括对客户更完整的分析和理解。Clementine 使你的企业在多方面受益。

- 可以改善客户获得和保持。
- 提高客户的生命周期价值。
- 识别并最小化风险和欺诈。
- 缩短产品开发过程中质量维护的周期。
- 支持科学研究。

作为一个数据挖掘平台，Clementine 结合商业技术可以快速建立预测性模型，进而应用到商业活动中，帮助人们改进决策过程。强大的数据挖掘功能和显著的投资回报率使得 Clementine 在业界久负盛誉。同那些仅仅着重于模型的外在表现，而忽略了数据挖掘在整个业务流程中的应用价值的其他数据挖掘工具相比，Clementine 其功能强大的数据挖掘算法，使数据挖掘贯穿于业务流程的始终，在缩短投资回报周期的同时极大地提高了投资回报率。

为解决各种商务问题，企业需要以不同的方式来处理各种类型迥异的数据，相异的任务类型和数据类型就要求有不同的分析技术。Clementine 提供最出色、最广泛的数据挖掘技术，确保可用最恰当的分析技术来处理相应的问题，从而得到最优的结果以应对随时出现的商业问题。即便改进业务的机会被庞杂的数据表格所掩盖，Clementine 也能最大限度地执行标准的数据挖掘流程，找到解决商业问题的最佳答案。

其次，为了推广数据挖掘技术，以解决越来越多的商业问题，SPSS 和一个从事数据挖掘研究的全球性企业联盟制定了关于数据挖掘技术的行业标准——CRISP-DM。与以往仅仅局限在技术层面上的数据挖掘方法论不同，CRISP-DM 把数据挖掘看做一个商业过程，并将其具体的商业目标映射为数据挖掘目标。最近一次调查显示，50% 以上的数据挖掘工具采用的都是 CRISP-DM 的数据挖掘流程，它已经成为事实上的行业标准。

Clementine 在数据挖掘流程的每一个环节中都支持 CRISP-DM 这一行业标准。这可以帮助你的公司把注意力集中在使用数据挖掘解决业务问题上,而不是集中在为每个项目发明一个新流程上。Clementine 提供的 CRISP-DM 项目管理器可以帮助你有效管理每个项目。这不但规避了许多常规错误,而且其显著的智能预测模型有助于快速解决出现的问题。

CRISP-DM 将数据挖掘技术与具体商业目标相结合,使得数据挖掘成为一个商业过程;这一过程始于对商业目标深入理解,终于数据挖掘结果的有效配置,如图 2-2 所示。

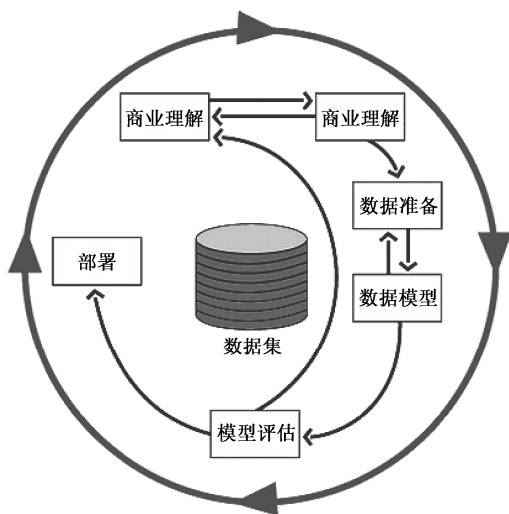


图 2-2 CRISP-DM 过程图

CRISP-DM 模型为一个 KDD 工程提供了一个完整的过程描述。该模型将一个 KDD 工程分为 6 个不同的,但顺序并非完全不变的阶段。综合上几节中的介绍,各个阶段的具体功能如下。

- 商业理解 (Business Understanding), 在第一个阶段必须从商业的角度了解项目的要求和最终目的是什么,并将这些目的与数据挖掘的定义,以及结果结合起来。
- 数据的理解 (Data Understanding) 及收集,对可用的数据进行评估。
- 数据的准备 (Data Preparation),对可用的原始数据进行一系列的组织的清洗,使之达到建模需求。
- 应用数据挖掘工具建立模型 (Modeling)。
- 对建立的模型进行评估 (Evaluation),重点具体考虑得出的结果是否符合第一步的商业目的。
- 部署 (Deployment),即将其发现的结果,以及过程组织成为可读文本形式(数据挖掘报告)。

在数据挖掘项目中使用 Clementine 应用模板 (CATs) 可以获得更优化的结果。应用模板完全遵循 CRISP-DM 标准,借鉴了大量真实的数据挖掘实践经验,是经过理论和实践证明的有效技术,为项目的正确实施提供了强有力的支撑。Clementine 中的应用模板包括以下几个方面。

- CRM CAT: 针对客户的获取和增长,提高反馈率并减少客户流失。
- Web CAT: 单击顺序分析和访问行为分析。

- Telco CAT：客户保持和增加交叉销售。
- Crime CAT：犯罪分析及其特征描述，确定事故高发区，联合研究相关犯罪行为。
- Fraud CAT：发现金融交易和索赔中的欺诈和异常行为。
- Microarray CAT：研究和疾病相关的基因序列并找到治愈手段。

2.3 SPSS 数据挖掘的过程

2.2 节中提到的数据挖掘的 CRISP-DM 程序模型，为数据挖掘项目的生命周期提供了一个综合的描绘。它包括了一个数据挖掘项目所要经历的各个阶段，各阶段的任务，以及这些任务之间的关系。从本质上来说，这些任务之间是否存在关系，取决于使用者的目的、背景及其利益所在，与此同时，更重要的还在于数据。

数据挖掘项目的周期由 6 个阶段组成。从图 2-2 中可以看出数据挖掘过程的各个阶段，这些阶段之间的顺序并不是固定不变的。在接下来的内容中，将利用一个数据挖掘的案例来简要说明图 2-2 中每一个阶段的轮廓。

此案例讨论的是关于产品市场定位及其市场细分的研究及应用。此数据挖掘的主要目的是 ABC 公司利用新产品赢得市场竞争地位，对现在的市场进行细分，以吸引目标客户，提高销量。这里严格按照 CRISP-DM 过程来进行此项目的数据挖掘。

2.3.1 商业理解

此阶段主要是对项目目标的理解，以及从商业角度考虑，对客户需求的理解，进而把这些理解转化为一个数据挖掘的定义和为达到目标的初步方案。所以，对应于 ABC 公司的商业目标，为了赢得市场竞争地位，决定推出新产品 Magic，该种产品的目标客户是中产阶层。为了进一步了解这种人群的心理特征、定位自己的产品、吸引目标客户，ABC 公司进行市场调研，具体从 8 个因子方面来进行调研，具体即消费因子、时尚因子、社会因子、爱国因子、期望因子、偏好因子、个性因子，以及家庭因子，从而了解这些目标客户的心理特征。然后根据数据挖掘，确定目标客户。

2.3.2 数据理解

数据理解阶段开始于数据的收集工作。研究者使用九点量表测量 400 名被试者对 30 项陈述的态度，标准变量是通过调研询问被试者对 Magic 型汽车的态度来测量，标准变量的测量通过九点量表来测试消费者对“我愿意购买 ABC 公司生产的 Magic 型汽车”的态度。接下来就是熟悉数据的工作，例如，检查数据的质量，对数据有初步的理解等。通过数据可以把消费者根据不同的特征分为几类，然后进一步了解不同特征的客户对 ABC 公司推出的新产品 Magic 的态度。最后可以根据不同的消费者制定不同的营销策略，以提高销量。

2.3.3 数据准备

数据准备阶段主要是将原始数据集转换为最终的进行数据挖掘的数据集，数据的准备工作有可能被多重实施。此过程中要对原始数据进行制表、整理、数据变量的选择、转换

等,必要时为了适应建模工具而进行数据清洗等。数据准备工作在整个数据挖掘流程所耗费的精力和时间是最多的,所以此阶段的工作尤为重要。

2.3.4 数据模型

根据数据理解阶段的初步判断,要对消费者进行分类,显然可以利用统计学中的聚类分析来实现此数据挖掘过程。因此,本项目的目的是通过聚类分析,将原始变量分别聚类成3类和4类,比较两种方法的效果。

2.3.5 评估

从数据分析的角度考虑,此阶段中,已经建立了一个或者多个高质量的模型。但是在进行最终的模型部署之前,更加彻底的评估模型,回顾构建模型过程中执行的每一个步骤是非常必要的,这样可以确保这些模型是否可以达到企业的目标。

首先,对数据集进行聚类分析,将样本聚类为3类或者4类。聚类为3类后的组重心参见表2-1。

表 2-1 聚类为 3 类后的组重心

| 因子 | 1 类 | 2 类 | 3 类 |
|------|---------|---------|---------|
| 消费因子 | -.45298 | .16364 | .29950 |
| 时尚因子 | .36038 | -.22794 | -.15239 |
| 社会因子 | .28739 | -.32881 | .00765 |
| 爱国因子 | .25444 | .70915 | -.87203 |
| 期望因子 | .52946 | -.29355 | -.26021 |
| 偏好因子 | .18363 | .11953 | -.28471 |
| 个性因子 | .00228 | .20936 | -.18616 |
| 家庭因子 | .56772 | -.64844 | .01414 |

聚类为3类后的方差分析参见表2-2。

表 2-2 聚类为 3 类后的方差分析表

| 因子 | Cluster | | Error | | f | Sig. |
|------|-------------|----|-------------|-----|---------|------|
| | Mean Square | df | Mean Square | df | | |
| 消费因子 | 21.981 | 2 | .894 | 397 | 24.579 | .000 |
| 时尚因子 | 13.717 | 2 | .936 | 397 | 14.656 | .000 |
| 社会因子 | 12.311 | 2 | .943 | 397 | 13.054 | .000 |
| 爱国因子 | 88.593 | 2 | .559 | 397 | 158.563 | .000 |
| 期望因子 | 29.241 | 2 | .858 | 397 | 34.092 | .000 |
| 偏好因子 | 8.863 | 2 | .960 | 397 | 9.228 | .000 |
| 个性因子 | 5.122 | 2 | .979 | 397 | 5.231 | .006 |
| 家庭因子 | 47.951 | 2 | .763 | 397 | 62.806 | .000 |

聚类为 4 类后的组重心参见表 2-3。

表 2-3 聚类为 4 类后的组重心

| 因子 | 聚类 | | | |
|------|---------|---------|---------|---------|
| | 1 类 | 2 类 | 3 类 | 4 类 |
| 消费因子 | .24279 | .01208 | .07032 | -.25561 |
| 时尚因子 | -.62391 | -.20361 | 1.03981 | -.22257 |
| 社会因子 | -.38800 | -.36427 | -.19066 | .77448 |
| 爱国因子 | .07062 | .46403 | -.43228 | -.09292 |
| 期望因子 | -.13134 | -.17016 | -.44699 | .62549 |
| 偏好因子 | .32674 | -.21434 | .01774 | -.07861 |
| 个性因子 | .73262 | -.29600 | .02012 | -.32019 |
| 家庭因子 | -.63160 | .97956 | -.26867 | -.14187 |

聚类为 4 类后的方差分析参见表 2-4。从表中的分析结果可以看出，聚类为 3 类或者 4 类时 8 个因子的组间差异都很显著。但是从方差分析的结果看，聚类为 4 类的结果不如聚类为 3 类的效果好。因此，比较可见，应该采用根据公因子得分进行聚类分析，最佳的类数是 3 类。

表 2-4 聚类为 4 类后的方差分析

| 因子 | Cluster | | Error | | f | Sig. |
|------|-------------|----|-------------|-----|--------|------|
| | Mean Square | df | Mean Square | df | | |
| 消费因子 | 4.398 | 3 | .974 | 396 | 4.515 | .004 |
| 时尚因子 | 49.658 | 3 | .631 | 396 | 78.651 | .000 |
| 社会因子 | 33.007 | 3 | .758 | 396 | 43.573 | .000 |
| 爱国因子 | 13.697 | 3 | .904 | 396 | 15.154 | .000 |
| 期望因子 | 22.929 | 3 | .834 | 396 | 27.497 | .000 |
| 偏好因子 | 4.910 | 3 | .970 | 396 | 5.060 | .002 |
| 个性因子 | 22.608 | 3 | .836 | 396 | 27.033 | .000 |
| 家庭因子 | 46.792 | 3 | .653 | 396 | 71.647 | .000 |

2.3.6 部署

模型的创建并不是项目的最终阶段，尽管建模是为了增加更多关于数据的信息，但是这些信息仍然需要以一种客户能够使用的方式被组织和呈现。在很多的案例中，部署阶段往往是客户而不是数据分析师，然而，对客户而言，预先了解需要执行的活动从而正确地使用已经构建的模型是非常重要的。

对于此案例，根据聚类结果，可以把消费者分为三类。

- 年轻创业型：经济状况不是很好，消费态度比较谨慎，追求时尚，较为关注社会问题，比较爱国，对将来充满自信，预期乐观，生活态度总的来说比较保守，个性比较平和稳重，看重家庭和婚姻生活。

- 中产稳健型：经济状况小康，适当消费，不追求时尚，不大关注社会问题，非常爱国，对将来的预期比较保守，较容易尝试新事物，较为注重享受和生活质量，自信，在周围人中有较强的影响力，不太关注家庭生活。
- 保守低调型：消费观念较强，不太追求时尚，对社会问题关注较少，国家观念淡薄，对将来的预期比较低调，生活方式保守，不太愿意尝试新事物，个性中庸，家庭观念一般。

也可以根据列联表来进行分析，可以了解三个类型的消费者对 ABC 公司的新产品 Magic 型汽车的态度。

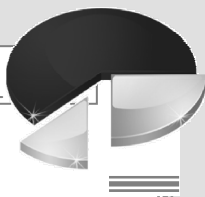
最后就可以给出此产品的营销建议。

通过对公因子进行的聚类分析，将 ABC 公司的目标客户分成了三个类型，这三种类型的消费者各自具有自己的独特特征。ABC 公司应该针对他们不同的特征和消费心理制定不同的营销策略。

年轻创业型的消费者对将来预期乐观，有奋斗精神，他们有较强的社会和家庭责任感。目前经济情况一般，消费态度较为谨慎。这部分人对 Magic 型汽车的态度最为友好，是公司主要的目标客户群。同时，这部分人极具成长潜力。公司应该针对这部分人的经济情况和消费心理，推出时尚创新、价格适中的汽车，广告的诉求上应该针对这部分人的心理特征，强调社会和家庭责任感。同时，公司应该关注这部分人的成长，尽力吸引其顾客忠诚度，因为将来这部分人进入中年，经济状况改善，有可能成为 Magic 型公司高档轿车的主要消费群。

中产稳健型的消费者对 Magic 型公司汽车的态度较好。公司应该针对这部分人的需求，推出注重舒适和享受，价格较高，质量高档的轿车。在广告诉求和产品宣传上，应该强调爱国的因素，从情感和经济两方面打动消费者。

保守低调型消费者对 Magic 型公司汽车的态度较为不好。这部分人不是公司主要的目标客户，但是也不能忽视，因为他们在总的消费群中占的比重相当大。公司应该加强对这部分客户的宣传和交流，提供关于公司产品的更多的信息，强调 Magic 型公司汽车稳健和高质量的特征，以吸引这部分消费者。



第3章 数据文件、变量与函数

数据文件的管理是进行数据分析之前所必须进行的过程,是数据分析的基础,主要内容有 SPSS 的变量设置、函数应用,以及数据的文件新建、编辑、打开、保存等。本章主要介绍 SPSS 中的变量类型、函数和数据文件的打开和保存。



本讲内容

- SPSS 的变量类型
- 数据文件的打开和保存
- SPSS 的函数

3.1 SPSS 的变量类型

在变量视图中,选择变量类型一列,然后单击“类型”按钮,弹出“定义变量类型(Define Variable Type)”对话框,如图 3-1 所示,用户可根据具体资料的属性对数据进行格式化。“变量类型”对话框中列出如下 9 种数据类型。

- 数值型,同时定义数值的宽度,即整数部分+小数点+小数部分的位数,默认为 8 位;定义小数位数,默认为 2 位。
- 加显逗号的数值型,即整数部分每 3 位数加一逗号,其余定义方式同数值型。
- 3 位加点数值型,无论数值大小,均以整数形式显示,每 3 位加一小点(但不是小数点),可定义小数位置,但都显示 0,且小数点用逗号表示。如 1.2345 显示为 1.234,500(实际是 $12345E-4$)。
- 科学记数法型,同时定义数值宽度和小数位数,在数据管理窗口中以指数形式显示。例如,定义数值宽度为 9,小数位数为 2,则 345.678 显示为 $3.4678E+02$ 。
- 日期型,用户可从系统提供的日期显示形式中选择自己需要的。例如,选择 mm/dd/yy 形式,则 2015 年 6 月 25 日显示为 06/25/15。

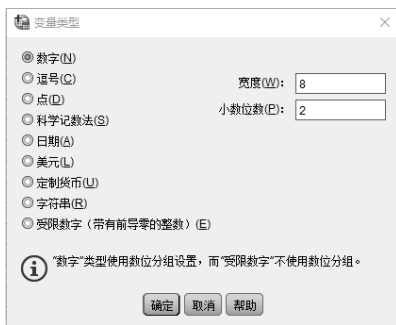


图 3-1 “变量类型”对话框

- 美元型，显示形式为数值前有 \$。
- 定制货币型，用户可从系统提供的形式中选择，并定义数值宽度和小数位数。常用的显示为整数部分每 3 位加一逗号，用户可定义数值宽度和小数位数。如 12345.678 显示为 12,345.678。
- 字符串型，用户可定义字符长度（Characters）以便输入字符。用户选择完毕可单击 Continue 按钮返回定义变量类型对话框。
- 受限数字。

3.1.1 数据的输入

定义好变量并格式化数据之后，即可向数据管理窗口输入原始数据。数据管理窗口的主要部分就是电子表格，横方向为电子表格的行，其行头以 1,2,3,... 表示，即第 1,2,3,... 行；纵方向为电子表格的列，其列头以 var00001,var00002,var00003,... 表示变量名。行列交叉处称为单元格，即保存数据的空格。鼠标一旦移入电子表格内即呈十字形，这时单击鼠标左键可激活单元格，被激活的单元格以加粗的边框显示；用户也可以按方向键上下左右移动来激活单元格。单元格被激活后，用户即可向其中输入新数据或修改已有的数据。如图 3-2 所示为一个已输入数据的数据管理窗口。为方便起见，用户也可省略定义变量和数据格式化两个步骤，一启动 SPSS 即向数据管理窗口中输入原始数据，这时，变量名默认为 var00001,var00002,var00003 等。




图 3-2 数据管理窗口

1. 插入一个新观测量

在菜单“数据（Data） 合并文件”上单击添加个案命令，可以在光标所在位置的上行插入一行新的观测个体，可以输入新的观测数据。

2. 查找指定的观测量

在数据窗口单击按钮 ，弹出一个对话框，如图 3-3 所示，输入要找的观测量的序号

后，单击“跳转”按钮，数据表中光标就会指到选定的观测量个体。

3.1.2 变量的编辑

建立数据文件的第一步是定义变量。在数据编辑窗口左下角激活“变量视图”窗口，如图 3-4 所示。



图 3-3 “转到个案”对话框

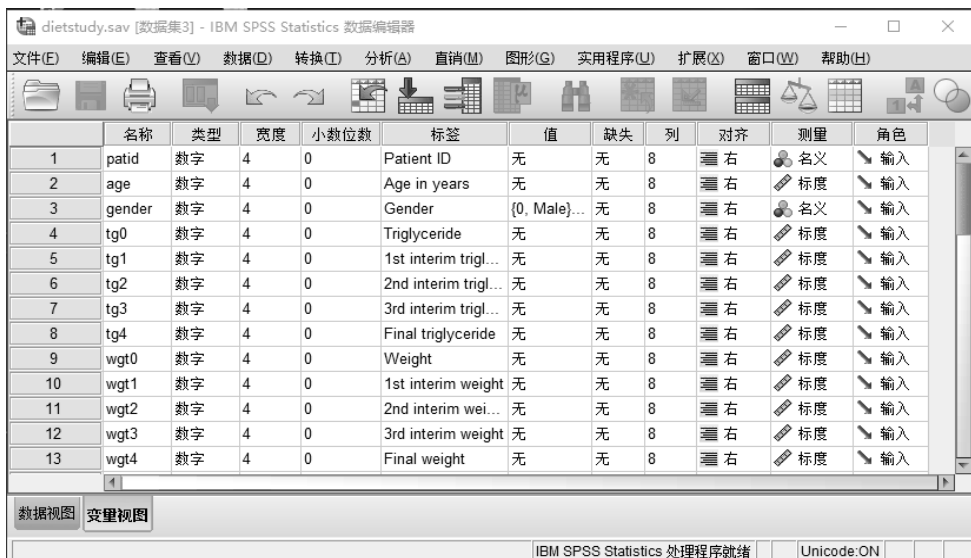


图 3-4 “变量视图”窗口

在数据窗口中，用户定义数据变量的名称、数据类型、宽度、小数位和标记等信息。

1. 名称选项

用于输入字符（汉字和英文）作为变量的名称，变量名应遵循下列原则。

- 每个变量名称必须是唯一的，不允许重复。
- 变量名称最多可包含 64 个字节，第一个字符必须是字母或字符 @、# 或 \$。
- 应避免用句点或下画线结束变量名称。
- 变量名不能与 SPSS 的关键字相同，即不能用 ALL、AND、BY、EQ、GT、LE 等。

2. 类型选项

用于设置变量类型，前面已经讲述，单击其左侧的按钮即可弹出对话框进行设置。

3. 宽度选项

用于指定数据字符占据的总个数（包括小数点和小数位）。

4. 小数位数选项

用于指定小数位数。

5. 标签选项

变量标签：有的时候变量名不能正确反映变量含义，有必要给它贴上标签以便识别。这个时候，就在变量定义的标签栏里输入注释。

6. 值选项

变量值标签：变量值标签是用来帮助解释某些变量，特别是分类变量的数值含义。例如，有一个数值变量，0 表示女性，1 表示男性。此时，为了便于识别这些数值，则用变量值标签。

7. 缺失选项

缺失值：缺失值是统计分析时，对数据中缺少数据的一种统计识别值。缺失值定义窗口如图 3-5 所示。

- 没有缺失值 (No missing values)，系统默认值用圆点 “.” 表示。
- 离散缺失值 (Discrete missing values)，可以定义三个缺失值。例如，第一格输入 “0”，表示凡为 0 的数据是缺失值。
- 范围加上一个可选离散缺失值 (Range plus one optional discrete missing value)，例如，低 (Low) 为 1，高 (High) 为 5，离散值 (Discrete value) 为 10，表示 1 ~ 5 的数据及数值 10 视为缺失值。



图 3-5 “缺失值 (Missing)” 对话框

8. 列选项

数据列的显示宽度：显示数据的列宽，默认 8 个字符。

9. 对齐选项

对齐方式：有左、中、右三种数据显示方式。

10. 测量选项

可以将测量级别指定为刻度 (Scale)、有序 (Ordinal) 或名义 (Nominal)。该选项仅用于统计绘图时坐标轴变量的区分，以及决策树模块的变量定义。定量变量，如虫口数、死亡率等；等级变量，如防治效果的好、一般、不好等；定性变量，如害虫抗药性发生低抗、中抗和高抗等。

11. 角色选项

角色包括输入、目标，两者都是无分区拆分。

3.2 数据文件的打开和保存

和大多数应用软件相同，SPSS 中数据文件的管理功能基本上都集中在了文件菜单上，该菜单的组织结构和其他软件也极为相似。

3.2.1 打开 SPSS 数据文件

从菜单选择“文件 打开 数据”命令，在弹出的“打开文件”对话框中指定数据文件的路径，文件名框内显示的是 SPSS 数据文件，系统默认的文件类型为“*.sav”，单击所选文件，单击“打开”按钮，或双击所选文件。这样就把该数据文件调入数据编辑窗口中。如图 3-6 所示，打开数据集 adl.sav。



图 3-6 打开 SPSS 数据集

3.2.2 打开其他格式的数据文件


SPSS 现在可以直接读入许多格式的数据文件，其中就包括 Excel 各个版本的数据文件。选择菜单“文件 打开 数据”或直接单击快捷工具栏上的按钮，系统就会弹出“打开文件”对话框，单击“文件类型”列表框，在里面能看到直接打开的数据文件格式，参见表 3-1。

表 3-1 数据文件格式

| 数 据 格 式 | 说 明 |
|-----------------------|-----------------------|
| SPSS (*.sav) | SPSS 数据文件 |
| SPSS 压缩文件 (*.zsav) | 被压缩的 SPSS 数据文件 |
| SPSS/PC+ (*.sys) | SPSS 4.0 版数据文件 |
| Systat (*.syd) | *.syd 格式的 Systat 数据文件 |
| Systat (*.sys) | *.sys 格式的 Systat 数据文件 |
| SPSS portable (*.por) | SPSS 便携格式的数据文件 |
| Excel (*.xls) | Excel 数据文件 |
| Lotus (*.w*) | Lotus 数据文件 |
| SYLK (*.slk) | SYLK 数据文件 |
| dBASE (*.dbf) | dBASE 系列数据文件 |
| dBASE (*.dbf) | dBASE 系列数据文件 |
| Text (*.txt) | 纯文本格式的数据文件 |
| data (*.dat) | 纯文本格式的数据文件 |
| SAS | SAS 格式的数据集 |

选择所需的文件类型,然后选中需要打开的文件,SPSS 就会按要求打开要使用的数据文件,并自动转换为数据 SPSS 格式。

除了上述直接打开数据集以外,也可以使用文本导入向导读入文本文件选择菜单“文件 打开文本数据”。系统就会弹出“打开文件”对话框,和前面的情况完全一样,只是文件类型自动跳到了 Text (*.txt)。

可以直接打开的数据文件格式也可通过打开数据集来观察,如图 3-7 所示。

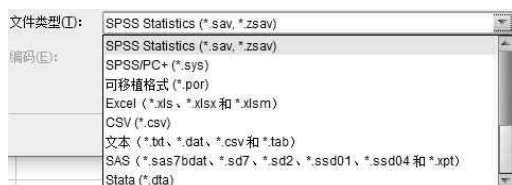


图 3-7 其他数据格式

3.2.3 数据文件保存

在对数据做了修改后,保存数据文件是必不可少的工作之一。选择菜单“文件 保存”,如果数据文件曾经存储过,则系统会自动按原文件名保存数据;否则,就会弹出和选择“另存为”菜单时相同的“另存为”对话框。里面可以保存的数据类型和可以打开的几乎一样多,选择合适的类型,单击“确定”按钮。

SPSS 可以将数据保存为 SPSS (*.sav)、Excel (*.xls)、dBASE (*.dbf)、ASCII (*.dat,*.txt) 等数据文件形式。在弹出的“保存文件”对话框里,指定保存路径,输入文件名,确定数据类型,最后单击“保存”按钮,如图 3-8 所示。



图 3-8 “保存文件”对话框

3.3 SPSS 函数

SPSS 函数大致可分为 8 类:算术函数、统计函数、分布函数、逻辑函数、缺失值函数、字符串函数、日期函数和其他函数。

函数的表示方法:函数的一般表达方式是在函数关键字后面括号中写入函数自变量。

函数自变量：函数自变量可以是单值或变量名，以及算术表达式的形式。如果使用变量名或带有变量名的表达式作为自变量，则必须在使用该函数之前对这些变量赋值，使函数类型为数值型。

下面将重点介绍算术函数和统计函数，并对一些常用的 SPSS 函数给出一般性的解释。

3.3.1 算术函数

算术函数是最常用的函数，可以满足对变量进行一般的运算，算术函数主要参见表 3-2。

表 3-2 算术函数

| 函 数 名 | 自变量含义 | 函 数 类 型 | 函数功能及说明 |
|--------------------------|---------------------|---------|--------------------|
| ABS (numexpr) | (算术表达式) * | 数值型函数 | 求绝对值 |
| ARSIN (numexpr) | (角度；弧度单位) | 数值型函数 | 求反正弦值 |
| ARTAN (numexpr) | (角度；弧度单位) | 数值型函数 | 求反正切值 |
| COS (radians) | (角度；弧度单位) | 数值型函数 | 求余弦值 |
| EXP (numexpr) | (算术表达式) | 数值型函数 | 求 e 的指数幂值 |
| LGI0 (numexpr) | (算术表达式) | 数值型函数 | 求以 10 为底的对数值 |
| LN (numexpr) | (算术表达式) | 数值型函数 | 求以 e 为底的对数 |
| MOD (numexpr, modulus) | (算术表达式；模数 (常数)) | 数值型函数 | 求算术表达式除以模数的余数 |
| SIN (radians) | (角度；弧度单位) | 数值型函数 | 求正弦值 |
| SQRT (numexpr) | (正数) | 数值型函数 | 求平方根 |
| RND (numexpr) | (算术表达式) | 数值型函数 | 求算术表达式的值四舍五入后的整数 |
| TRUNC (numexpr) | (算术表达式) | 数值型函数 | 求算术表达式的值被截去小数部分的整数 |

3.3.2 统计函数

统计函数也是统计分析中常用的函数之一，主要反映变量的数据特征，时间序列的滞后变量等，具体函数参见表 3-3。

表 3-3 统计函数

| 函 数 名 | 自变量含义 | 函 数 类 型 | 函数功能与说明 |
|---------------------------------|-------------------|-------------|--|
| CFVAR (numexpr, numexpr, ...) | (变量名, 变量名, ...) | 数值型函数 | 求出多个变量值的变异系数 (标准差/均值) |
| LAG (variable) | (变量名) | 数值型函数或字符型函数 | 返回滞后一期的变量数据。对第一个观测量来说，将返回系统缺失值，如果指定的变量是字符型，则返回空格 |
| LAG (variable, ncases) | (变量名, 自然数 n) | 数值型函数 | 返回滞后 n 期的变量数据。对前 n 个观测量来说，将返回系统缺失值，如果指定的变量是字符型，则返回空格 |

续表

| 函 数 名 | 自变量含义 | 函 数 类 型 | 函数功能与说明 |
|-----------------------------------|-------------------|---------|---|
| MAX (ivalue,value[,...]) | (变量名, 变量名, ...) | 数值型函数 | 求多个变量值中的最大值; 例如, MAX (数学, 物理, 化学): 分别计算每个学生三门成绩中的最高分 |
| MEAN (numexpr,numexpr,...) | (变量名, 变量名, ...) | 数值型函数 | 求多个变量值的平均值; 例如, MEAN (数学, 物理, 化学): 分别计算每个学生三门成绩的平均值 |
| MIN (value,value[,...]) | (变量名, 变量名, ...) | 数值型函数 | 求多个变量值中的最小值; 例如, MIN (数学, 物理, 化学): 分别计算每个学生三门成绩中的最低分 |
| NVALID (variable, variable,...) | (变量名, 变量名, ...) | 数值型函数 | 求出变量的 (不包括缺失值) 数量 |
| SD (numexpr,numexpr,...) | (变量名, 变量名, ...) | 数值型函数 | 求多个变量值的标准差; 例如, SD (数学, 物理, 化学): 分别计算每个学生三门成绩的标准差 |
| SUM (numexpr,numexpr,...) | (变量名, 变量名, ...) | 数值型函数 | 求多个变量值的和; 例如, SUM (数学, 物理, 化学): 分别计算每个学生三门成绩的总和 |
| VARIANCE (numexpr,numexpr,...) | (变量名, 变量名, ...) | 数值型函数 | 求多个变量值的方差; 例如, VARIANCE (数学, 物理, 化学): 分别计算每个学生三门成绩的方差 |

3.3.3 逻辑函数

逻辑函数有如下两个。

1. ANY (test,valu,value,...) 逻辑型函数

自变量为 (变量名,x1,x2,...), 函数功能是判断变量值是否是 x1,x2,... 中的一个, 例如, Any (数学,80,90,70): 分别对每条个案判断其数学成绩是否为 80 分或 90 分或 70 分。

2. RANGE (test,lo,hi[, 10, hi...]) 逻辑型函数

变量必须都为数值型或字符型, 自变量为 (变量名,x1,x2), 其中 x1 ~ x2, 函数功能是判断某变量值是否 x1 ~ x2, 例如, RANGE (数学,80,90): 分别对每条个案判断其数学成绩是否为 80 ~ 90 分。

3.3.4 日期和时间函数

日期和时间函数如下所示。

1. DATE.DMY (day,month,year)

SPSS 日期型格式的数值函数, 返回与指定的日、月、年相应的日期值。要正确显示这个值, 必须将变量赋予 DATE 格式。自变量必须为整数。day 的范围为 1 ~ 31, month 的范围为 1 ~ 12, year 的范围在 4 位数时要大于 1582, 2 位数时应是该世纪的后两位年代数值。

2. DATE.YRDAY (year,daynum)

SPSS 日期型的格式数值函数, 返回与指定的天数、年相应的日期值。要正确显示这个值, 必须赋予其 DATE 格式。daynum 取值范围为 1 ~ 366。

3. XDATE.DATE (datevalue)

SPSS 日期格式的数值型函数，从具有 SPSS 的日期格式的自变量数值返回一个日期，自变量数值由 DATE.xxx 函数产生或按 DATE 输入格式读取。该函数用于将日期的数值格式转换为日期格式，因此，要想按日期格式显示必须再在 Variable View 中定义一种日期格式，否则会按 SPSS 日期的数值格式显示。

4. XDATE.HOUR (datevalue)

数值型函数，从 DATE.xxx 函数产生或按一种 DATE 格式读入的 SPSS 日期格式的数值，返回一个小时数（0~23）。

5. XDATE.JDAY (datevalue)

数值型函数，通过 DATE.xxx 函数产生或由 DATE 输入格式读入 SPSS 日期格式的数值，返回一年的天数（1~366）。

6. XDATE.MDAY (datevalue)

数值型函数，从一个 SPSS 日期格式的数值通过 DATE.xxx 函数产生或由 DATE 输入格式读入，返回一个月的天数（1~31）。

7. XDATE.MINUTE (datevalue)

数值型函数，通过 DATE.xxx 函数产生或由 DATE 输入格式读入 SPSS 日期格式的数值，返回分钟数（0~59）。

8. XDATE.MONTH (datevalue)

数值型函数，通过 DATE.xxx 函数产生或由 DATE 输入格式读入 SPSS 日期格式的数值，返回一年中的月数（1~12）。

9. XDATE.TDAY (timevalue)

数值型函数，自变量是由 TIME.xxx 函数产生或由 TIME 输入格式读取的 SPSS 时间间隔格式的数值，返回整天数（正整数）。

10. XDATE.TIME (datevalue)

SPSS 时间间隔格式的数值型函数，把自变量的值看作从午夜开始的秒数，返回一天中的时间（时、分、秒）。自变量是 SPSS 日期格式的数值，可以由 DATE.xxx 函数产生的或由 DATE 输入格式读入的。由该函数建立的变量应该给定一个合适的显示格式。在 Variable View 中，赋予它一个时间显示格式，将变量值显示成小时和分。

11. XDATE.WEEK (datevalue)

数值型函数。由一个 SPSS 日期格式数值（由 DATE.xxx 函数产生或由一种 DATE 输入格式读入），返回周数（1~53 整数）。

12. XDATE.WKDAY (datevalue)

数值型函数, 由一种通过 DATE.xxx 函数产生或用 DATE 格式读入的 SPSS 日期格式数值, 返回的数值表示一周的星期几(星期一至星期日用 1~7 之间的整数表示)。

13. XDATE.YEAR (datevalue)

数值型函数, 由 DATE.xxx 函数产生或用 DATE 格式读入的 SPSS 日期格式的数值, 返回年数。

YRMODA (year, month, day) 数值型函数, 返回一个由 1582 年 10 月 15 日到自变量给定的年月日 (year, month, day) 之间的天数。

3.3.5 随机变量函数

随机变量函数的一般形式为 RV.分布名(参数,...)。其中圆点前是函数类名, 圆点后是分布名称, 圆点是半角的圆点, 括号内是自变量。自变量是分布参数。如果在数据文件中建立新变量时使用这些函数, 变量值的个数等于数据文件中有效观测量数。函数值为产生服从指定统计分布的随机序列。下面列出常用的分布函数的随机数。

1. NORMAL (stddev)

数值型函数, 产生一个来自均值为 0 标准差为 stddev 的分布总体的随机数。

2. RV.BERNOULLI (P)

数值型函数, 产生一个来自伯努利分布具有指定概率参数 P 的随机数。

3. RV.BINOM (n, P)

数值型函数, 产生一个来自二项式分布具有指定试验次数 n 和概率参数 P 的随机数。

4. RV.CHISQ (df)

数值型函数, 产生一个来自卡方分布具有指定自由度 df 的随机数。

5. RV.EXP (shape)

数值型函数, 产生一个来自指数分布具有指定形状参数的随机数。

6. RV.F (df1, df2)

数值型函数, 产生一个来自 F 分布具有指定自由度的随机数。

7. RV.GEOM (p)

数值型函数, 产生一个来自几何分布具有指定概率参数 P 的随机数。

8. RV.HYPER (totd, sample, hits)

数值型函数, 产生一个来自超几何分布具有指定参数的随机数。

9. RV.LOGISTIC (mean, scale)

数值型函数，产生一个来自逻辑斯蒂分布具有指定的均数 mean 和标度 scale 参数的随机数。

10. RV.LNORMAL (a, b)

数值型函数，产生一个来自对数正态分布具有指定参数的随机数。

11. RV.NORMAL (mean, stddev)

数值型函数，产生一个来自正态分布具有指定均值 mean 和标准差 stddev 的随机数。

12. RV.PARETO (threshold, shape)

数值型函数，产生一个来自帕累托分布具有指定临界值 threshold 和形状 shape 参数的随机数。

13. RV.POISSON (mean)

数值型函数，产生一个来自泊松分布具有指定均值或比率参数的随机数。

14. RV.T (df)

数值型函数，产生一个来自学生 t 分布具有指定自由度的随机数。

15. RV.UNIFORM (min, max)

数值型函数，产生一个来自具有指定最大值 max 和最小值 min 的均匀一致分布的随机数。

16. RV.WEIBULL (a, b)

数值型函数，产生一个来自威布尔分布具有指定参数的随机数。

17. UNIFORM (max)

数值型函数，产生一个来自一致分布的值在 0 和自变量给定的 max 之间的伪随机数。自变量 max 必须是一个数值，但可以是负数。

3.3.6 反分布函数

反分布函数的一般形式为 $\text{IDF.分布名}(P, \text{参数}, \dots)$ 。其中圆点前是函数类名，圆点后是分布名称，括号内是自变量。第一个自变量 P 是这个分布的累积概率，其后的自变量是指定分布的参数。函数值是相应分布的累计概率值为 P 的临界值。

$\text{IDF.CHISQ}(p, df)$ 数值型函数，产生来自卡方 χ^2 分布的临界值，第一个自变量为概率值 P ，第二个自变量为自由度 df 。例如，累积概率为 0.95，自由度为 5 的卡方分布的临界值记作 $\text{IDF.CHISQ}(0.95, 5)$ ，其函数值 $\text{IDF.CHISQ}(0.95, 5) = 1.145$ 。

$\text{IDF.EXP}(p, \text{shape})$ 数值型函数，产生一个来自指数分布的临界值，该分布具有给

定行状参数 $shape$, 概率值为 P 。

IDF.F ($p, df1, df2$) 数值型函数, 产生一个来自 F 分布的值, 该分布具有自由度为 $df1$, $df2$, 累计概率为 P 的临界值。例如, 显著性概率在 0.05 水平上, 自由度分别为 6、5 的 F 值为 $IDF.F(0.95, 6, 5) = 4.9503$ 。

DF.LOGISTIC ($prob, mean, scale$) 数值型函数, 产生一个均值为 $mean$ 和标度参数为 $scale$, 累计概率为 P 的逻辑斯蒂分布的临界值。

IDF.LNORMAL (P, a, b) 数值型函数, 产生具有指定参数和累计概率为 P 的对数正态分布的临界值。

IDF.NORMAL ($p, mean, stddev$) 数值型函数, 产生来自正态分布具有指定均值和标准差的累计概率。例如, 显著性水平为 0.05 , 均值为 0 , 标准差为 1 的标准正态分布的临界值 $IDF.NORMAL(0.95, 0, 1) = 1.645$ 。

IDF.PARETO ($prob, threshold, shape$) 数值型函数, 产生一个来自帕累托分布, 累计概率为 P 的值, 该分布的临界值为 $threshold$, 尺度参数为 $scale$ 。

IDF.T ($prob, df$) 数值型函数, 产生一个自由度为 df , 累计概率为 P 的来自学生 t 分布的临界值。

IDF.UNIFORM (P, min, max) 数值型函数, 产生一个累计概率为 P 的来自均匀分布的临界值, 均匀分布的最大值为 max 、最小值为 min 。

PROBIT (P) 数值型函数, 产生累计概率为 P 的标准正态分布的临界值。

3.3.7 累计分布函数

累计分布函数的一般形式为 CDF.分布名 ($q, 参数, \dots$) , 其中圆点前是函数类名, 圆点后是分布名称, 括号内是自变量。第一个自变量 q 是符合分布的数值, 后面的自变量是相应分布的参数。函数值是相应分布的随机变量取值小于等于 q 的概率值。

1. CDF.BERNOULLI (q, P)

数值型函数, 产生来自具有给定概率参数 P 的伯努利分布, 变量值小于 q 的累计概率值。

2. CDF.BETA ($q, shape1, shape2$)

数值型函数, 产生来自 Bate 分布的变量取值小于 q 的累计概率值, 该分布具有给定的形状参数 $shape1$, $shape2$ 。

3. CDF.BINOM (q, n, P)

数值型函数, 产生来自二项分布的变量取值小于 q 的累计概率值, 该分布具有给定每次试验成功的概率 P , 成功的试验次数是 n 。当 $n=1$ 时, 该函数与 CDF.BERNOULLI 相同。

4. CDF.CAUCHY ($q, loc, scale$)

数值型函数, 产生来自柯西分布的变量取值小于 q 的累计概率值, 该分布具有给定的位置参数 loc 和标度参数 $scale$ 。

5. CDF.CHISQ (q , df)

数值型函数，返回来自卡方分布的变量取值小于 q 的累计概率值，该分布具有给定的自由度 df 。

6. CDF.EXP (q , $shape$)

数值型函数，产生来自指数分布的变量取值小于 q 的累计概率，该分布具有给定的形状参数 $shape$ 。

7. CDF.F (q , $df1$, $df2$)

数值型函数，产生来自 F 分布的变量取值小于 q 的累计概率值，该分布具有给定的自由度 $df1$, $df2$ ，累计概率值小于 $quant$ 。

8. CDF.GAMMA (q , $shape$, $scale$)

数值型函数，产生来自伽玛分布的变量取值小于 q 的累计概率，该分布具有给定的形状参数 $shape$ 和标度参数 $scale$ 。

9. CDF.GEOM (q , P)

数值型函数，产生一个几何分布的变量取值小于 q 的累积概率，即获得一次成功的试验次数，成功概率由 P 确定。

10. CDF.HYPER (q , $total$, $sample$, $hits$)

数值型函数，产生小于 q 的累积概率，即具有指定特性的事件数 q ，当样品 $sample$ 事件是从尺寸为 $total$ 的总体中随机选择出来的情况下，其命中数 $hits$ 具有指定的特性。

11. CDF.LAPLACE (q , $mean$, $scale$)

数值型函数，产生来自拉普拉斯分布的变量取值小于 q 的累计概率，该分布具有给定的均值 $mean$ 和标度参数 $scale$ 。

12. CDF.LOGISTIC (q , $mean$, $scale$)

数值型函数，产生来自逻辑斯蒂分布的变量取值小于 q 的累计概率，该分布具有给定的均值 $mean$ 和标度参数 $scale$ 。

13. CDF.LNORMAL (q , a , b)

数值型函数，产生具有指定参数的对数正态分布变量取值小于 q 的累计概率值。

14. CDF.NEGBIN (q , $thresh$, P)

数值型函数，产生变量取值小于 q 的累计概率值，即当临界参数为 $thresh$ ， P 给出成功的概率。

15. CDFNORM (zvalue)

数值型函数，产生一个具有均值为 0，标准差为 1 的随机变量的取值小于 zvalue 的概率。

16. CDF.NORMAL (q, mean, stddev)

数值型函数，产生一个正态分布的变量取值小于 q 的累计概率，该分布均值为 mean，标准差为 stddev。

17. CDF.PARETO (q, threshold, shape)

数值型函数，产生一个变量取值小于 q 的帕累托分布的累计概率，该分布具有指定的限值 threshold 和形状参数 shape。

18. CDF.POISSON (q, mean)

数值型函数，产生一个来自 POISSON 分布的小于 q 的累计概率值，它具有指定的均值或率参数。

19. CDF.T (q, df)

数值型函数，产生一个变量取值小于 q 的学生 t 分布的累计概率，该分布具有指定的自由度参数 df。

20. CDF.UNIFORM (q, min, max)

数值型函数，产生一个变量取值小于 q 的均匀一致分布的累计概率，该分布具有指定的最小值 min 和最大值 max 参数。

21. CDF.WEIBULL (q, a, b)

数值型函数，产生一个变量取值小于 q 的威布尔分布的累计概率，该分布具有指定的参数。

3.3.8 缺失值函数

缺失值函数有如下四种。

1. NMISS (variable, ...)

数值型函数，自变量是当前工作数据文件中的变量名。计算自变量中缺失值的数目。例如，Missing (数学)：分别对数学这个变量逐个判断是否为系统缺失值或用户缺失值。1 表示是，0 表示不是。

2. MISSING (variable)

逻辑型函数，自变量应该是工作数据文件中的变量名。如果变量具有缺失值，返回 1 或者 true。

3. SYSMIS (numvar)

逻辑型函数，自变量 numvar 是工作数据文件中的一个数值型变量的变量名。如果 numvar 的值为系统缺失值，返回 1 或者 true。

4. VALUE (variable)

数值型或字符型函数，忽略用户缺失值，即将用户缺失值看作普通的数据，返回变量值。自变量必须是工作数据文件中的变量名。

3.3.9 字符串函数

1. CONCAT (strexpr, strexpr, ...)

字符型函数，函数中每个自变量都是一个字符串表达式。该函数值是一个字符串，是各自变量代表的字符串按括号中的顺序串接起来的。此函数要求两个或两个以上的自变量。

2. INDEX (haystack, needle)

数值型函数，产生一个整数，它表明字符串 needle 在字符串 haystack 中第一次出现的起始位置。如果返回值为 0，表明字符串 needle 不在字符串 haystack 中存在。例如，INDEX ("ABCDEFGH", "DE")：找到字符串 DE 在字符串 ABCDEFGH 中第一次出现的位置，INDEX ("ABCDEFGH", "DE")=4。

3. INDEX (haystack, needle, divisor)

数值型函数，见 INDEX (haystack, needle) 函数。其第三个自变量 divisor 是可选择的，它必须是一个整数，表明将字符串 needle 均匀地分为要被查询的独立字符串的字符数。

4. LENGTH (strexpr)

数值型函数，自变量是字符串，函数值是字符串表达式值的长度。这里获得的长度包括尾部空格。

5. LPAD (strexpr, length)

字符型函数，第一个自变量 strexpr 是字符串，第二个自变量 length 是正整数，其范围为 1~255。函数值是字符串表达式的左侧增加空格扩展到 length 所规定的长度。

6. LTRIM (strexpr)

字符型函数，返回的字符串是自变量表达式的值去除打头的空格后的字符串。

7. LOWER (strexpr)

字符型函数，返回字符串，将字符串中的大写字母改变为小写字母。

8. RINDEX (haystack, needle)

数值型函数，产生一个整数，它表明字符串 needle 在字符串 haystack 中最后出现的开

始位置。返回 0 表示字符串 needle 不在 haystack 中。

9. RPAD (strexpr, length)

字符型函数，返回字符串，其长度由 length 决定。在字符串表达式的右侧加空格，以达到 length 的长度，length 的值为 1 ~ 255。

10. RPAD (strexpr, length, char)

字符型函数，返回字符串，见 RPAD (strexpr, length) 函数。第三个变量 char 是可以选择使用的，它表示在字符串的右侧增加一个字符 char。char 必须是一个带有引号的单个字符或其值是单个字符的字符表达式。

11. RTRIM (strexpr)

字符型函数，返回截取了尾部空格后的字符串。该函数通常用于大字符串表达式中，要把压缩了尾部空格的字符串赋予一个变量。

12. RTRIM (strexpr, char)

字符型函数，返回截取了尾部字符 char 后的字符串。char 必须是一个带有引号的单个字符或其值是单个字符的字符表达式。

13. STRING (numexpr, format)

字符型函数，根据 format 所设定的格式将数值表达式转换为字符串。例如，string (-1.5,F5.2) 返回字符串 '-1.50'。第二个自变量 format 必须是写一个数值的格式。

14. SUBSTR (strexpr, pos)

字符型函数，返回字符串表达式中从 pos 开始到其结尾处的子字符串。

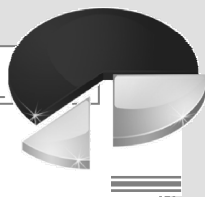
15. SUBSTR (strexpr, pos, length)

字符型函数，返回字符串表达式中从 pos 开始长度为 length 的子字符串。

16. UPCAS (strexpr)

字符型函数，返回将字符串表达式中小写字符变为大写字符串。

注意：数值与数字有区别，以上所讲的数值是数，数字指的是表现为数字的字符。



第4章 数据预处理

在 SPSS 中,数据文件的编辑、整理等功能被集中在了 Data 和 Transform 两个菜单项中,这两个菜单主要是在进行数据分析之前,进行数据的初步预分析。本章将详细介绍数据文件的整理和数据变量的变换、计算。



本讲内容

- 数据文件的整理
- 数据变量的变换和计算

4.1 数据文件的整理

在许多情况下,先要对数据进行一些整理(如分组、合并、加权等)才能将其用于最终的统计分析。这些功能基本上都集中在数据(Data)菜单项中,各个子菜单的功能参见表 4-1。下面对一些特别的对话框做详细介绍。

表 4-1 Data 菜单项详细功能

| 名 称 | 功 能 |
|---|--|
| 定义变量属性 (Define Variable Properties) | 定义变量值的标签 |
| 复制变量属性 (Copy Data Properties) | 以一个外部 SPSS 文件为模板为活动数据集定义其中选定变量的变量值标签或者数据集的性质 |
| 定义日期 (Define Dates) | 自动生成时间变量 |
| 定义多重相应集 (Define Multiple Response Sets) | 定义复选变量 |
| 验证 (Validation) | 定义单个变量或者交叉变量的有效性,以及检验变量值的有效性 |
| 标识重复个案 (Identify Duplicate Cases) | 识别重复的观测量 |
| 标识异常个案 (Identify Unusual Cases) | 识别不寻常的观测量 |
| 个案排序 (Sort Cases) | 观测量排序 |
| 排序变量 (Sort Variables) | 变量值排序 |
| 转置 (Transpose) | 数据文件转置 |
| 重组 (Restructure) | 数据格式重排 |

续表

| 名 称 | 功 能 |
|--------------------------|---------------------------------------|
| 合并文件 (Merge Files) | 数据文件合并 |
| 汇总 (Aggregate) | 数据分类汇总 |
| 正交设计 (Orthogonal Design) | 用于自动生成正交设计表格, 用于统计分析中做正交设计 |
| 复制数据集 (Copy Dataset) | 为当前工作数据文件复制一个新的数据集作为模板, 用于定义文件或者变量的性质 |
| 拆分文件 (Split File) | 数据文件拆分 |
| 选择个案 (Select Cases) | 观测量选择 |
| 个案加权 (Weight Cases) | 观测量加权 |

4.1.1 个案排序 (Sort Case) 过程

个案排序过程用于对数据集的变量进行排序。选择菜单“数据 (Data) 个案排序 (Sort Case)”命令, 则弹出如图 4-1 所示的对话框, 如果要对某变量进行排序, 则选中后选入“排序依据 (Sort by)”变量框中即可。

1. 排序依据 (Sort by) 选项栏

选入需要进行排序的变量。

2. 排列顺序 (Sort Order) 选项栏

用于设置排序方法。包括升序 (Ascending) 和降序 (Descending) 排列。

例如, 对数据集 stroke_valid.sav 中的变量 age 进行排序。选中变量 age 到“排序依据 (Sort by)”选项栏, 并选择“升序 (Ascending)”选项栏, 然后单击“确定”按钮进行排序, 结果如图 4-2 所示。



图 4-1 个案排序设置对话框

| | hospid | hospsize | patid | physid | age |
|---|--------|----------|--------------|--------|-----|
| 1 | NHV | | 1 6531754596 | 371884 | 45 |
| 2 | OZN | | 3 4182500237 | 883285 | 45 |
| 3 | NSR | | 1 2404065713 | 037350 | 45 |
| 4 | NHV | | 1 5182610884 | 190855 | 46 |
| 5 | QWS | | 1 3116422532 | 817329 | 46 |
| 6 | RLD | | 1 5811956549 | 560175 | 46 |

图 4-2 排序结果

4.1.2 转置 (Transpose) 过程

转置过程用于对数据文件中的行列进行转换。选择菜单“数据 (Data) 转置 (Transpose)”, 弹出如图 4-3 所示的对话框。



图 4-3 “转置 (Transpose) 设置”对话框

1. 变量 (Variable(s)) 选项栏

变量框，将要进行转置的变量移入其中。

2. 名称变量 (Name Variable) 选项栏

变量命名框。在原变量列表框中选择一个变量移入这个框内，这个变量就作为转置后的新变量名，一般默认情况下，系统将自动生成新变量 Var001, Var002 等变量名。

例如，对数据集 Workprog.sav 进行转置操作。选中变量到“变量 (Variable(s))”选项栏中，单击图 4-3 中的“确定”按钮进行转置。结果如图 4-4 所示。

| | CASE_LBL | var001 | var002 | var003 | var004 | var005 | var006 | var007 |
|---|----------|--------|--------|--------|--------|--------|--------|--------|
| 1 | age | 16.00 | 17.00 | 17.00 | 19.00 | 18.00 | 17.00 | 17.00 |
| 2 | marital | .00 | .00 | .00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 3 | incbef | 8.00 | 8.00 | 8.00 | 9.00 | 7.00 | 8.00 | 8.00 |
| 4 | incaft | 12.00 | 10.00 | 11.00 | 18.00 | 12.00 | 15.00 | 13.00 |
| 5 | ed | 1.00 | 2.00 | 1.00 | 1.00 | 1.00 | 1.00 | 2.00 |
| 6 | gender | | | | | | | |
| 7 | reside | 1.00 | 1.00 | 1.00 | 3.00 | 3.00 | 2.00 | 3.00 |
| 8 | prog | .00 | .00 | .00 | 1.00 | .00 | 1.00 | .00 |

图 4-4 数据集转置结果

4.1.3 合并文件 (Merge File) 过程

合并文件过程用于数据文件的合并，可以添加新的变量，也可以添加观测量。此过程分为两个菜单，即添加个案 (Add Cases) 和添加变量 (Add Variables)。

1. 添加个案

添加个案添加观测量，选择菜单“数据 (Data) 合并文件 (Merge File) 添加个案 (Add Cases)”，则弹出如图 4-5 所示对话框，如果数据集已经打开，则就会显示在“打开的数据集 (An Open Dataset)”选项框中，否则可以单击“外部 SPSS Statistics 数据文件 (An external SPSS data file)”选项框旁边的“浏览”按钮打开数据集，选好后，则单击“继续 (Continue)”按钮。

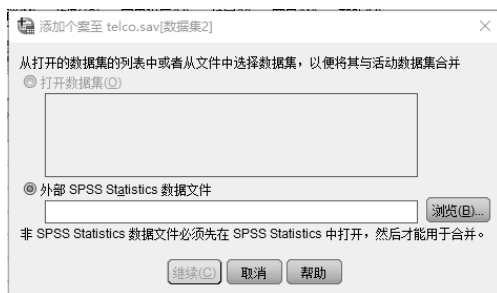


图 4-5 “添加个案 (Add Cases) 设置”对话框

单击“继续 (Continue)”按钮以后则弹出如图 4-6 所示的“添加个案”对话框。图中各组成部分详细功能如下所述。

- 非成对变量 (Unpaired Variables): 其中列出了两个文件中所有不同的变量名。
- 新的活动数据集中的变量 (Variables in New Active Dataset): 其中列出了新文件所有的变量名, 可以通过按钮将变量移入到非成对变量框 (Unpaired Variables) 中。
- 指示个案源变量 (Indicate case source as variable): 如果选择此项, 则系统将在新文件中创建一个新变量, 用来标记观测量来自哪个文件, 系统默认为 source01。



图 4-6 “添加个案”对话框

2. 添加变量

添加变量添加观测量, 选择菜单“数据 (Data) 合并文件 (Merge File) 添加变量 (Add Variables)”。则弹出如图 4-7 所示的“添加变量”对话框。同如图 4-5 所示对话框一样, 选好数据后, 则单击“继续 (Continue)”按钮即可继续设置。



图 4-7 “添加变量 (Add Variables)”对话框

单击“继续 (Continue)”按钮以后则弹出如图 4-8 所示的“添加变量”对话框。



图 4-8 “添加变量 (Add Variables From)”对话框

图中各组成部分详细功能如下所述。

- 排除的变量 (Exclude Variables)。
- 新的活动数据集 (New Active Dataset)。
- 按键变量匹配个案 (Match cases on key variables in sorted file): 当两个文件含有数目不同的观测值时选择这一项进行合并, 其中包含三个选项。两个文件都提供个案 (Both Files Provide) 表示合并关键变量值相等的观测值为一个变量。非活动数据集为基于关键字的表 (Non-active dataset is keyed table) 表示仅合并两个文件关键变量值相等的观测量到新文件中, 且只保留活动数据集中所有的观测量。活动数据集为基于关键字的表 (Active dataset id keyed table) 表示仅合并两个文件关键变量值相等的观测量到新文件中, 且只保留非活动数据集中的所有的观测量。
- 键变量 (Key Variables): 其中放置将两个文件联系起来的关键变量。
- 指示个案源变量 (Indicate case source as variable): 如果选择这个选项, 则系统将在新文件中创建一个新的变量, 用来标记观测量是来自哪个文件的, 且系统默认的变量名为 source01。

4.1.4 汇总 (Aggregate) 过程

汇总过程用于数据分类汇总操作, 选择菜单“数据 (Data) 汇总 (Aggregate)”, 弹出如图 4-9 所示的“汇总设置”对话框, 组成部分如下所述。

1. 分界变量 (Break Variable(s)) 选项栏

选入分界变量。



图 4-9 “汇总 (Aggregate) 设置”对话框

2. 汇总变量 (Summaries of Variable (s)) 选项栏

其中选入的是待汇总的变量，单击其下的“函数 (Function)”按钮，则弹出如图 4-10 所示的“汇总函数”对话框，每个选项代表了一个汇总函数。如果要改变系统默认的新变量名，则单击“变量名与标签 (Name & Label)”按钮，在弹出的对话框中修改即可。

- 摘要统计量 (Summary Statistics)
- 特定值 (Specific Values)
- 个案数 (Number of Cases)
- 百分比 (Percentages)
- 分数 (Fractions)
- 计数 (Counts)



图 4-10 “汇总函数”对话框

3. 保存 (Save) 选项栏

其下有三个选项，各选项功能如下。

- 将汇总变量添加到活动数据集 (Add aggregated variables to active dataset) 中。
- 创建只包含汇总变量的新数据集 (Create a new dataset containing only the aggregated variables)
- 写入只包含汇总变量的新数据文件 (Write a new data file containing only the aggregated variables)

4. 用于大型数据集的选项 (Options for Very Large Datasets) 选项栏

此选项栏是为大型数据集提供的。

- 文件已经按分界变量排序 (File is already sorted on break variable (s))
- 汇总前对文件进行排序 (Sort file before aggregating)

4.1.5 拆分文件 (Split File) 过程

拆分文件过程用于数据文件的拆分, 数据文件的拆分指的是将数据按照某个或者几个变量分成一些供统计分析的分组。

选择菜单“数据 (Data) 拆分文件 (Split File)”, 则弹出如图 4-11 所示的“拆分文件设置”对话框, 各组成部分如下。

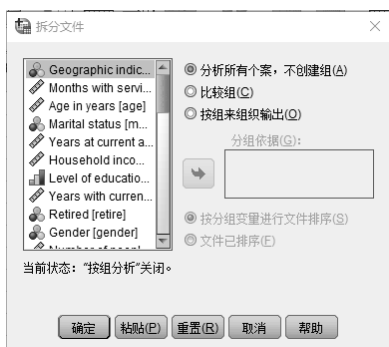


图 4-11 “拆分文件 (Split File) 设置”对话框

分析所有个案, 不创建组 (Analyze all cases, do not create groups)

比较组 (Compare Groups), 若选择这一项, 在进行了统计分析后, 分界变量将安置在同一个表格中比较输出。

按组来组织输出 (Organize Output by Groups)

按分组变量进行文件排序 (Sort the file by grouping variables)

文件已排序 (File is already sorted)

4.1.6 选择个案 (Select Cases) 过程

选择个案过程用于从指定数据文件中选取符合要求的观测量作为样本参与数据分析。

选择菜单“数据 (Data) 选择个案 (Select Cases)”, 则弹出如图 4-12 所示的“选择个案设置”对话框, 各组成部分如下。

1. 输出 (Output) 选项栏

用于设置输出的对话框, 有三个选项。

- 过滤掉未选定的个案 (Filter out Unselected Cases), 即未被选中的观测量仍然保留在数据文件中。
- 将选定个案复制到新数据集 (Copy selected cases to a new dataset) 中, 并在下面的数据集名称 (Dataset Name) 栏里命名这个数据集。
- 删除未选定的个案 (Delete Unselected Cases)

2. 选择 (Select) 选项栏

关于观测量选择的单选框。

- 所有个案 (All Cases)

- 如果条件满足 (If Condition Satisfied) : 选择满足条件的观测量。
- 随机个案样本 (Random Sample of Cases) : 随机抽取观测量样本。
- 基于时间或个案范围 (Bases on Time or Case Range) : 按照时间或者观测量范围选择。
- 使用过滤变量 (Use Filter Variable) : 使用器变量选择观测量。



图 4-12 “选择个案 (Select Cases) 设置”对话框

4.1.7 个案加权 (Weight Cases) 过程

个案加权过程用于对观测量进行加权操作。选择菜单“数据 (Data) 个案加权 (Weight Cases)”，则弹出如图 4-13 所示的“个案加权设置”对话框，各组成部分详细如下。

选项栏个案加权 (Weight Cases by) 用于设定全变量，选中后则激活其下的选项框，将这个变量选入此选项框中就可。

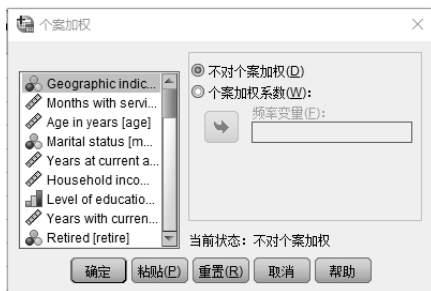


图 4-13 “个案加权 (Weight Cases) 设置”对话框

4.2 数据变量的变换和计算

数据变量的变换和计算需要利用转换 (Transform) 菜单来实现，此菜单中有很多子菜单，其下的各个菜单的功能详细参见表 4-2。下面主要对一些菜单选项进行介绍。

表 4-2 转换菜单功能简介

| 名 称 | 功 能 |
|---|------------------|
| 计算变量 (Compute Variable) | 根据原始数据计算而得的新变量数据 |
| 对个案内的值计数 (Count Values Within Cases) | 变量值计数功能 |
| 重新编码为相同变量 (Recode into Same Variables) | 变量的重新赋值 1 |
| 重新编码为不同变量 (Recode into Different Variables) | 变量的重新赋值 2 |
| 自动重新编码 (Automatic Recode) | 自动重新赋值功能 |
| 可是离散化 (Visual Binning) | 变量组段划分 |
| 最优离散化 (Optimal Binning) | 最优组段划分 |
| 个案排序 (Rank Cases) | 变量值排序 |
| 日期和时间向导 (Date and Time Wizard) | 日期时间向导功能 |
| 创建时间序列 (Create Time Series) | 创建时间序列 |
| 替换缺失值 (Replace Missing Values) | 缺失值替代功能 |
| 随机数字生成器 (Random Number Generator) | 随机数生成器 |

4.2.1 计算变量 (Compute Variables) 过程

计算变量过程用于根据原始数据计算而得到变量数据。选择菜单“转换 (Transform) 计算变量 (Compute Variables)”，则弹出如图 4-14 所示的“计算变量”对话框，各组成部分如下所述。



图 4-14 “计算变量 (Compute Variable)”对话框

1. 目标变量 (Target Variable) 选项栏

用于定义将要产生的目标变量,在空白处填入目标变量的名称,可以是一个新变量名,也可以是已经定义的变量名,单击其下的“类型与标签 (Type & Label)”按钮,则出现如图 4-15 所示的“类型与标签”对话框,在对话框中可以定义目标变量的标签 (Label) 和类型 (Type)。



图 4-15 “类型与标签”对话框

2. 数字表达式 (Numeric Expression) 选项栏

用于填写数字表达式,放置新变量的计算表达式。

3. 计算器板

类似于计算器界面的框,其中的按钮就是 SPSS 的所有运算符。

4. 函数组 (Function Group) 列表框和函数和特殊变量 (Functions and Special Variables) 选项栏

此栏是函数组列表框和函数与特殊变量框。函数组列表框里列举了 SPSS 中的所有函数组,单击一个组名,这一组的所有函数和特殊变量将出现在函数和特殊变量选项栏中,再单击任意一个函数名,这个函数的信息就会出现在计算板下面的空白框中,供用户查询这个函数的意义和用法。

5. 如果 (If) 按钮

单击“如果 (If)”按钮,则弹出如图 4-16 所示的“如果设置”对话框,此对话框与图 4-14 所示对话框基本相同。



图 4-16 “如果 (If) 设置”对话框

4.2.2 计数（Count）过程

在统计分析中，有一项特定变量值计数功能，可以计数在一个观测量中满足特定要求的那些变量值出现的次数，并将结果记录在一个新变量中。

选择菜单“转换（Transform） 计算个案中值的出现次数（Count Values Within Cases）”，则弹出如图 4-17 所示的“计算个案中值的出现次数”对话框。

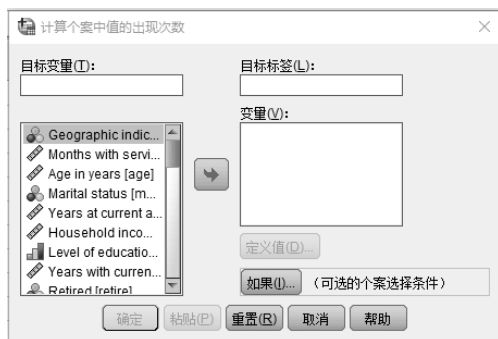


图 4-17 “计算个案中值的出现次数（Count Values Within Cases）”对话框

1. 目标变量（Target Variable）和目标标签（Target Label）选项栏

表示目标变量栏和目标变量标签栏，在目标变量栏内输入目标变量名，以保存计数结果；在目标标签选项栏内输入目标变量的标签。

2. 变量（Variables）选项栏

用于选入变量，输入将对其进行特定变量值计数的变量。单击其下的“定义值（Define Values）”按钮，弹出如图 4-18 所示的“定义值”对话框。



图 4-18 “定义值（Define Values）”对话框

- 值（Value）选项栏：其中有 6 个选项。值表示按照变量的指定值计数，在空白栏中填入这个指定值。

- 要计数的值 (Values to Count): 计数值框, 在值单选框选定后, 单击“添加 (Add)”按钮, 这个值或者范围就加入到其下的“要统计的值 (Values to Count)”选项框中, 可以单击“更改”按钮修改或者单击“删除”按钮删除即可。

3. 如果 (If) 按钮

用于选择观测量, 和上面所述功能一样, 此处不再累述。

4.2.3 重新编码 (Recode) 过程

重新编码过程用于变量的重新赋值, 选择菜单“转换 (Transform) 重新编码为相同变量 (Recode into Same Variables)”, 则弹出如图 4-19 所示的“重新编码为相同变量”对话框。如果选择此命令, 则系统会产生新变量值直接替代原始变量值。选择菜单“转换 (Transform) 重新编码为不同变量 (Recode into Different Variables)”, 则弹出如图 4-20 所示的“重新编码为不同变量”对话框。如果选择此命令, 系统将为产生的新变量值赋予一个新的变量。重新编码为不同变量对话框比重编码为相同变量对话框多一个输出变量 (Output Variable) 选项框。下面介绍重新编码为不同变量对话框。



图 4-19 “重新编码为相同变量 (Recode into Same Variables)”对话框



图 4-20 “重新编码为不同变量 (Recode into different Variables)”对话框

1. 输入变量 (Input Variable) → 输出变量 (Output Variable) 选项框

输入变量 输出变量选项框。此框用于显示将要重新赋值的原变量名和将要建立的

新变量名。

2. 输出变量选项框

输出变量选项框，用于定义新变量的名称和标签，输入完成以后单击“更改（Change）”按钮，新变量名出现在“输入变量 输出变量”选项框中。

3. 旧值和新值（Old and New Values）选项栏

单击“旧值”和“新值”按钮，则弹出如图 4-21 所示的“旧值和新值”对话框。



图 4-21 “旧值和新值（Old and New Values）”对话框

- 旧值（Old Value）：原变量值框。用于选择将要赋予新值的变量值。
- 新值（New Value）：新变量框，用于选择将赋予的新变量值。值表示直接为新变量赋予指定值；系统缺失（System-missing）表示为新变量赋予系统缺失值；复制旧值（Copy Old Value(s)）表示将原变量值直接赋予新变量值。
- 旧 新（Old New）选项栏：原变量值 新变量值框，用于显示由原变量值转化为新变量值的详细信息。
- 输出变量是字符串（Output variables are strings）：新变量值赋予字符型变量选项，选择此项，无论原始变量值是否为字符型都将被赋予为字符型变量。
- 将数字字符串转换为数字（Convert numeric strings to numbers）：表示将以数值作为字符串的字符型变量转换为数值型变量。

4. 如果（If）按钮

用于使用条件表达式对所选个案子集的值进行重新编码。

4.2.4 个案排秩（Rank Cases）过程

个案排秩过程用于变量值排序，即变量值求秩。选择菜单“转换（Transform） 个案排秩（Rank Cases）”，则弹出如图 4-22 所示的“个案排秩”对话框。



图 4-22 “个案排秩 (Rank Cases)”对话框

1. 将秩 1 指定给 (Assign Rank 1 to) 选项栏

用于将秩 1 赋给变量值。

- 最小值 (Smallest Value): 表示将秩 1 赋给最小变量值。
- 最大值 (Largest Value): 用于将秩 1 赋给最大变量值。

2. 变量 (Variable(s)) 选项栏

用于将要产生的新的秩变量的原文件中的变量选入其中, 新变量名的名字就是原变量名字前加字母“r”。

3. 依据 (By) 选项栏

系统将排序标准选项栏内的变量对观测值排序求秩, 如果不设定排序标准变量, 则系统会有多对观测值求秩。

4. 绑定值 (Ties)

单击图 4-22 中的“绑定值 (Ties)”按钮, 则弹出如图 4-23 所示的“结”对话框, 主要用于处理相等的观测值的秩的问题。

- 平均值 (Mean): 系统默认取排序后相同观测值处各秩次序的均值为相同观测值处的秩。
- 低 (Low): 取排序后相同观测值处各秩次序的最小值为相同观测值处的秩。
- 高 (High): 取排序后相同观测值处各秩次序的最大值为相同观测值处的秩。
- 顺序秩到唯一值 (Sequential ranks to unique values): 取排序时相同观测值处第一个出现的秩次值为相同观测值处的秩。

5. 秩的类型 (Rank Type) 按钮

此框用于选择秩的类型, 单击“秩的类型”按钮则弹出如图 4-24 所示的“秩的类型”对话框。

6. 显示摘要表 (Display Summary Tables) 选项

该选项为系统默认选项。

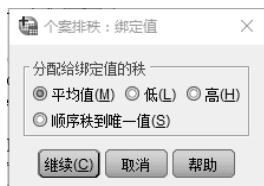


图 4-23 “结 (Ties)” 对话框



图 4-24 “秩的类型 (Rank Type)” 对话框

4.2.5 自动重新编码 (Automatic Recode) 过程

自动重新编码过程用于自动重新赋值功能，主要用于将数值型和字符型变量自动地转换成连续整数，重新编码 (Recode) 过程中需要自己设置每个原变量值转化为新变量值的转化方法，而自动重新编码过程则由系统按照一定的原则自动为原变量赋值。

选择菜单“转换 (Transform) 自动重新编码 (Automatic Recode)”，则弹出如图 4-25 所示的“自动重新编码”对话框。组成部分如下。



图 4-25 “自动重新编码 (Automatic Recode)” 对话框

1. 变量 (Variable) → 新名称 (New Name) 选项栏

其中列出所有被选中的旧变量名称，及其将转化为新变量名称。

2. 新名称 (New Name) 选项栏

新变量名栏，在其中输入新变量的名称，然后单击下面的“添加新名称 (Add New Name)”按钮，则这个新变量名将加入到变量 新名称选项栏中。

3. 重新编码的起点 (Recode Starting form) 选项栏

选择对变量的自动重新赋值是从最小值开始还是从最大值开始。

- 最低值 (Lowest Value): 从最小值开始。
- 最高值 (Highest Value): 从最大值开始。

4. 对所有变量使用相同的重新编码方案 (Use the same recording scheme for all variables) 选项

将一个变量的重新赋值方案给所欲选中的变量, 这里要求被选中的所有变量都是同一类型的。

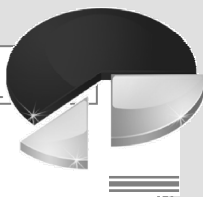
5. 将空字符串视为用户缺失值 (Treat blank string value as user-missing) 选项

用于当被选变量为字符型变量时, 将缺失值定义为用户缺失值, 这样这个缺失值不会被重新赋予一个整数, 而是用户缺失值, 若不选, 则系统将缺失值赋予一个整数。

6. 模板 (Template) 选项栏

此栏是“模板”选项框, 用于将当前文件中自动重新赋值方案作为一个模板保存在指定的文件中或者引用指定文件的模板到当前文件中。

- 从文件应用模板 (Apply Template from File): 引用模板。
- 将模板另存为 (Save Template as): 保存模板。



第5章 基本统计分析

统计科学分为两大部分：描述性统计和推断性统计。描述性统计提供了将原始数据整理成有用形式的方法，这些方法包括收集、整理、概括、描述及给出数据信息。具体来说，这些方法包括将统计资料整理成表格的形式，将统计资料整理成图形化的形式，用平均数、中位数、众数等度量集中趋势，用极差、标准差、变异系数等度量离散趋势。



本讲内容

- 基本概念
- 频数分析
- 描述性统计分析过程
- 数据探索性分析过程
- 交叉表分析过程

5.1 基本概念

在数据分析的时候，一般首先要对数据进行描述性统计分析（Descriptive Analysis），以发现其内在的规律，再选择进一步分析的方法。描述性统计分析要对调查总体所有变量的有关数据做统计性描述，主要包括数据的频数分析、数据的集中趋势分析、数据离散程度分析、数据的分布，以及一些基本的统计图形。

5.1.1 基本的统计概念

在介绍 SPSS 的描述性统计分析之前，先了解基本的统计概念。

1. 总体与样本

(1) 总体与个体

研究对象的全体称为总体。一般地把我们关心的随机变量 X 称为总体。组成总体的每个单元称为个体。个体也可理解为总体 X 的取值。

(2) 简单随机抽样

为了使抽样具有充分的代表性, 所以要求:

每个个体被抽到的机会均等。

每次抽取是独立的(共抽取 n 次)。

这样的抽样称为简单随机抽样。通常的抽样都是无放回的, 当总体很大时, 可以满足独立性。

(3) 样本

在总体中抽取 n 个个体, 称为总体的一个样本, 记为 (X_1, X_2, \dots, X_n) , 其中每次抽样 $X_i (i=1, 2, \dots, n)$ 也都是随机变量(解释), 共 n 个随机变量, 加上括号, 表示样本是一个整体。

(4) 样本的容量

抽取的个体数 n 称为样本的容量。

(5) 独立同分布

每次抽取的 X_i 来自总体, 应该与总体 X 有相同的分布(概率密度相同), 所以, 样本是一组具有独立同分布的随机变量。

(6) 样本观察值(样本值)

样本的测试结果记为 (X_1, X_2, \dots, X_n) , 是一组数据, 在容易产生误会时, 大小写要分清, 尤其在作理论分析时, 一般都取大写, 作为随机变量处理。

2. 统计量

统计量是含有样本 X_1, X_2, \dots, X_n 的一个数学表达式, 并且式中不含未知参数, 因而在得到样本值后立即算出它的数值来。

在抽样之前, 统计量的值无法确定, 抽样测试之后, 可以观察到它的取值, 因此, 统计量是随机变量, 是由样本派生出来的随机变量。

三个重要的统计量如下。

$$\text{样本均值 } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i。$$

$$\text{样本方差 } S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2。$$

$$\text{样本标准差 } S = \sqrt{S^2}。$$

其中 \bar{X} 作为均值可以反映总体 X 的均值(不是等同), S^2 是数据与均值偏离值平方的平均, 体现样本的离散程度, 因而可以反映总体 X 的方差。

3. 抽样分布

统计量既然是随机变量, 当然有它的概率分布, 称为抽样分布。以下仅给出结论, 结论都是对正态总体而言的。

(1) 样本均值的分布

若总体 $X \sim N(\mu, \sigma^2)$, 则 $X_i \sim N(\mu, \sigma^2)$ (独立同分布), 于是作为线性函数。

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

特别地，标准化以后，得

$$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

(2) t 分布

当总体标准差 σ 未知时， U 不再是统计量，这时可用样本标准差 S 代替，但不再是正态分布，而是一种新的分布 $T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$ 称为服从于自由度 $n-1$ 的 t 分布。它的密度曲线与正态曲线相类似。

(3) χ^2 分布

为了将样本方差 S^2 和总体相比较、联系。构造出

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

称为服从于自由度为 $n-1$ 的 χ^2 分布。

U 、 T 、 χ^2 是继 \bar{X} 、 S^2 、 S 后第二轮复合而成的统计量，可以更有利于实际的应用。

5.1.2 描述性统计分析

1. 数据的频数分析

在数据的预处理部分，曾经提到利用频数分析和交叉频数分析来检验异常值。此外，频数分析也可以发现一些统计规律。例如，收入低的被调查者用户满意度比收入高的被调查者高，或者女性的用户满意度比男性低等。不过这些规律只是表面的特征，在后面的分析中还要经过检验。

2. 数据的集中趋势分析

数据的集中趋势分析是用来反映数据的一般水平，常用的指标有平均值、中位数和众数等。各指标的具体意义如下所述。

平均值：是衡量数据的中心位置的重要指标，反映了一些数据必然性的特点，包括算术平均值、加权算术平均值、调和平均值和几何平均值。

中位数：是另外一种反映数据的中心位置的指标，其确定方法是把所有数据以由小到大的顺序排列，位于中央的数据值就是中位数。

众数：是指在数据中发生频率最高的数据值。

如果各个数据之间的差异程度较小，用平均值就有较好的代表性；而如果数据之间的差异程度较大，特别是有个别的极端值的情况，用中位数或众数有较好的代表性。

3. 数据的离散程度分析

数据的离散程度分析主要是用来反映数据之间的差异程度，常用的指标有方差和标准

差。方差是标准差的平方，根据不同的数据类型有不同的计算方法。

4. 数据的分布

在统计分析中，通常要假设样本的分布属于正态分布，因此，需要用偏度和峰度两个指标来检查样本是否符合正态分布。偏度衡量的是样本分布的偏斜方向和程度；而峰度衡量的是样本分布曲线的尖峰程度。一般情况下，如果样本的偏度接近于 0，而峰度接近于 3，就可以判断总体的分布接近于正态分布。

5. 绘制统计图

用图形的形式来表达数据，比用文字表达更清晰、更简明。在 SPSS 软件里，可以很容易绘制各个变量的统计图形，包括条形图、饼图和折线图等。

5.2 频率分析

本节介绍频率分析（Frequencies）的过程，此过程可以输出详细的频数分布表，能处理多种变量类型，详细操作如下。

5.2.1 频率分析过程的操作界面

选择菜单“分析（Analyze） 描述统计（Descriptive Statistics） 频率（Frequencies）”，弹出如图 5-1 所示的“频率设置”对话框，该对话框主要组成部分如下。

1. 变量选择

图 5-1 的左边即为待分析的变量列表，变量选项栏用于选择要产生频数表的变量，可以同时选择多个变量，系统会分别处理。



图 5-1 “频率（Frequencies）设置”对话框

2. 显示频率表（Display Frequency Tables）选项栏

用于显示频数表。

3. 统计量 (Statistics) 设置

单击图 5-1 中的“统计量 (Statistics)”按钮则弹出如图 5-2 所示的“统计量”对话框，此对话框主要用于选择需要计算的统计量。

- 百分位值 (Percentile Values): 包括四分位数、分割点和百分位数选项。
- 集中趋势 (Central Tendency): 描述集中趋势的统计量。
- 离散 (Dispersion): 描述性离散趋势的统计量。标准差 (Std.Deviation)、方差 (Variance)、范围 (Range)、最小值、最大值、标准误差平均值 (S.E.mean)。
- 分布 (Distribution): 描述变量分布情况的统计量。包括正态分布的偏度 (Skewness) 和峰度 (Kurtosis)。偏度是描述数据分布偏斜方向和程度的度量，峰度是描述数据分布形态的陡缓程度。
- 值为组的中点 (Values are group midpoints): 用于标识分位点是否恰好是变量的某个取值。

4. 图表 (Chart) 设置

单击“图表 (Chart)”按钮，弹出如图 5-3 所示的“图表”对话框，此对话框主要用于定义图形和图形中的数据。



图 5-2 “统计量 (Statistics)”对话框



图 5-3 “图表 (Chart)”对话框

- 图表类型 (Chart Type): 用于定义绘制的统计图类型，包括无 (None)、条形图 (Bar Charts)、饼图 (Pie Charts)、直方图 (Histograms)。
- 图表值 (Chart Values): 用于选择条形图或者饼图上的数据是显示变量值的频率还是百分比。

5. 格式 (Format) 设置

单击“格式”(Format)按钮,弹出如图 5-4 所示的“格式设置”对话框,主要用于定义频数表的输出。

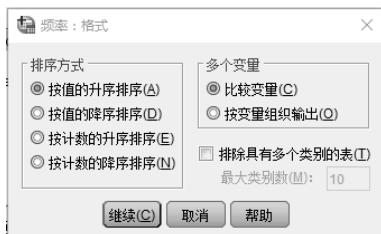


图 5-4 “格式 (Format) 设置”对话框

- 排序方式 (Order by): 用于定义频数表中的显示方式。
- 多个变量 (Multiple Variables): 用于定义当同时选择多个变量时是否在同一表中比较多个变量的统计量。比较变量 (Compare Variables) 表示将所有变量结果在一个图形中输出,以便比较。按变量组织输出 (Organize Output by Variables) 表示为每一个变量单独输出一个图形。
- 排除具有多个类别的表 (Suppress tables with more than n categories): 用于定义当频数表中个数大于 n 时不输出频数表。

5.2.2 实例分析

本实例所用数据集为 SPSS 自带的 contacts.sav 数据集,此数据集包含 5 个变量,共有 70 个观测量,数据集的格式如图 5-5 所示,下面就利用频数过程分析此数据集。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|------|----|----|------|--------------------|----------------|----|----|----|----|----|
| 1 | dept | 数字 | 4 | 0 | Department | {1, Develop... | 9 | 6 | 右 | 名义 | 输入 |
| 2 | rank | 数字 | 4 | 0 | Company rank | {1, Emplo... | 9 | 6 | 右 | 名义 | 输入 |
| 3 | sale | 数字 | 8 | 2 | Amount of last ... | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 4 | time | 数字 | 4 | 0 | Time since last... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 5 | size | 数字 | 4 | 0 | Size of company | {1, Very sm... | 无 | 6 | 右 | 名义 | 输入 |

图 5-5 数据集格式



结果文件

——附带光盘“PROGRAM\CH05\实例 5-1”文件夹



动画演示

——附带光盘“AVI\实例 5-1.avi”文件

1. 参数设置

选择菜单“分析 (Analyze) 描述统计 (Descriptive Statistics) 频率 (Frequencies)”。

弹出如图 5-6 所示的“频率设置”对话框，选择变量 dept 到“变量”选项栏中。

然后单击图 5-6 中的“图表”按钮，弹出如图 5-7 所示的“图表设置”对话框。选中直方图变量，并选择其下的“在直方图上显示正态曲线 (With Normal Curve)”选项，然后单击“继续 (Continue)”按钮。

2. 结果分析

设置完成后则单击“频率设置 (Frequencies Dialog Box)”对话框中的“确定 (OK)”按钮进行频数分析，结果如图 5-8 所示为频数分析结果，如 Computer services 变量的频数为 30，占有所有 dept 中的 42.9%，占有所知道的 dept 中的 48.4%。



图 5-6 “频率 (Frequencies) 设置”对话框



图 5-7 “图表 (Chart) 设置”对话框

| | | Department | | | |
|----|-------------------|------------|-------|-------|-------|
| | | 频率 | 百分比 | 有效百分比 | 累计百分比 |
| 有效 | Development | 16 | 22.9 | 25.8 | 25.8 |
| | Computer services | 30 | 42.9 | 48.4 | 74.2 |
| | Finance | 13 | 18.6 | 21.0 | 95.2 |
| | Other | 3 | 4.3 | 4.8 | 100.0 |
| | 总计 | 62 | 88.6 | 100.0 | |
| 缺失 | Don't know | 8 | 11.4 | | |
| 总计 | | 70 | 100.0 | | |

图 5-8 频数分析结果

图 5-9 所示为直方图，并绘制分布曲线。

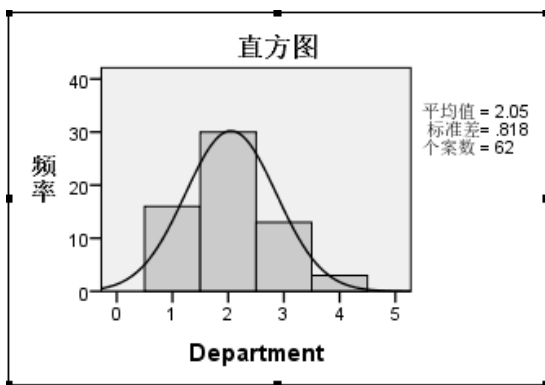


图 5-9 直方图

5.3 描述性统计分析过程

描述性统计分析过程是最基本的统计分析,对于数据集可以进行描述性统计分析,本节介绍 SPSS 描述性统计分析过程中经常出现的各类描述性统计量。

5.3.1 描述性统计分析过程参数设置

选择菜单“分析 (Analyze) 描述统计 (Descriptive Statistics) 描述 (Descriptive)”,弹出如图 5-10 所示的“描述”对话框,该对话框主要组成部分如下所述。

变量 (Variable(s)) 选项栏:用于选入要分析的变量,可以同时选择多个变量。

将标准化值另存为变量 (Save standardized values as variables):用于定义是否将原始数据的标准化结果保存在数据文件之中。

选项 (Options) 设置:单击图 5-10 中的“选项 (Options)”按钮,弹出如图 5-11 所示的“选项设置”对话框,此对话框主要用于选择需要计算的统计量。

- 平均值 (Mean) 和总和 (Sum):描述变量集中趋势的统计量。
- 离散 (Dispersion):描述变量离散程度的统计量。
- 分布 (Distribution):描述变量分布情况的统计量。
- 显示顺序 (Display Order):当选中多个变量时,选择变量分析结果的输出顺序。





图 5-10 “描述 (Descriptive)”对话框



图 5-11 “选项 (Options) 设置”对话框

5.3.2 实例分析

本实例所用数据集为 SPSS 自带的数据集 telco.sav,本数据集为某电信公司的用户调查数据,下面就利用描述性统计过程研究客户消费的变量 customer spending 中哪一种服务更被用户选择。数据集格式如图 5-12 所示。

-  **结果文件** —— 附带光盘 “PROGRAM\CH05\实例 5-2 ” 文件夹
-  **动画演示** —— 附带光盘 “AVI\实例 5-2.avi ” 文件

1. 参数设置

选择菜单 “分析 (Analyze) 描述统计 (Descriptive Statistics) 描述 (Descriptive)”, 弹出如图 5-13 所示的 “描述变量选择设置” 对话框, 选择变量 Long distance last month、Toll free last month、Equipment last month、Calling card last month, 以及 Wireless last month 到 “变量 (Variable(s))” 选项栏中。



| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|----|---------|----|----|------|--------------------|------------------|----|----|----|----|----|
| 1 | region | 数字 | 4 | 0 | Geographic indi... | {1, Zone 1}... | 无 | 6 | 右 | 名义 | 输入 |
| 2 | tenure | 数字 | 4 | 0 | Months with se... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 3 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 4 | marital | 数字 | 4 | 0 | Marital status | {0, Unmarrie... | 无 | 7 | 右 | 名义 | 输入 |
| 5 | address | 数字 | 4 | 0 | Years at curren... | 无 | 无 | 7 | 右 | 标度 | 输入 |
| 6 | income | 数字 | 8 | 2 | Household inco... | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 7 | ed | 数字 | 4 | 0 | Level of education | {1, Did not c... | 无 | 6 | 右 | 有序 | 输入 |
| 8 | employ | 数字 | 4 | 0 | Years with curr... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 9 | retire | 数字 | 8 | 2 | Retired | {00, No}... | 无 | 10 | 右 | 名义 | 输入 |
| 10 | gender | 数字 | 4 | 0 | Gender | {0, Male}... | 无 | 6 | 右 | 名义 | 输入 |

图 5-12 数据集格式



图 5-13 “描述 (Descriptive) 变量选择设置” 对话框

2. 结果分析

设置好变量后, 则单击 “确定 (OK)” 按钮进行描述性统计分析过程。结果如图 5-14 所示, 输出了观测量、最大最小值、均值和方差。如变量 Equipment last month 对应的 Meam (均值) 为 14.2198 和 Std.Deviation (标准差) 为 19.06854。

| 描述统计 | | | | | |
|--------------------------|------|-----|--------|---------|----------|
| | 个案数 | 最小值 | 最大值 | 平均值 | 标准差 |
| Long distance last month | 1000 | .90 | 99.95 | 11.7231 | 10.36349 |
| Toll free last month | 1000 | .00 | 173.00 | 13.2740 | 16.90212 |
| Equipment last month | 1000 | .00 | 77.70 | 14.2198 | 19.06854 |
| Calling card last month | 1000 | .00 | 109.25 | 13.7810 | 14.08450 |
| Wireless last month | 1000 | .00 | 111.95 | 11.5839 | 19.71943 |
| 有效个案数(成列) | 1000 | | | | |

图 5-14 描述性统计分析结果

5.4 数据探索性分析过程

数据探索性分析过程除了可以计算基本的统计量以外,也可以给出一些简单的检验结果和图形,有助于用户进一步地分析数据。

5.4.1 数据探索性分析过程参数设置

选择菜单“分析(Analyze) 描述统计(Descriptive Statistics) 探索(Explore)”,弹出如图 5-15 所示的“探索设置”对话框,该对话框主要组成部分如下。

变量列表(Dependent List)选项栏:用于选入待分析的变量,可以选择的各个变量。

因变量列表(Factor List)选项栏:用于选择分组变量,根据该变量取值不同,分组分析因子列表(Dependent List)中的变量。

个案标注依据(Label Cases by)选项栏:选择标签变量。

输出(Display)选项栏:用于定义输出结果,包括两者都(Both)、统计量(Statistics),以及图(Plots),系统默认为两者都(Both)。

统计量(Statistics)设置:单击图 5-15 中的“统计量(Statistics)”按钮,则弹出如图 5-16 所示的“统计量设置”对话框,此对话框用于设置各种需要计算的统计量。

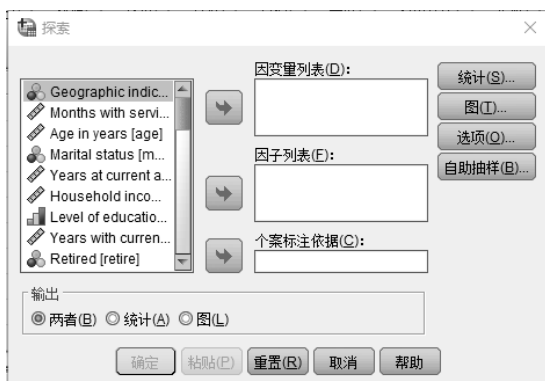


图 5-15 “探索(Explore)设置”对话框

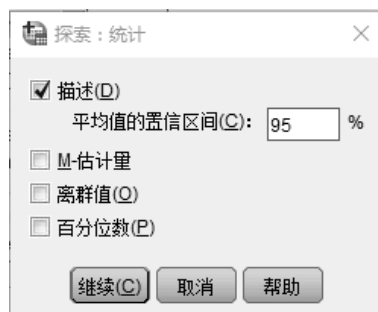


图 5-16 “统计量(Statistics)设置”对话框

- 描述(Descriptives): 计算一般的描述性统计量。输出平均数、中位数、众数、5%修正均数、标准误、方差、标准差、最小值、最大值、范围、四分位范围、峰度系数、峰度系数的标准误、偏度系数、偏度系数的标准误及指定的平均数可信区间。

- M-估计量 (M-estimators): 描述集中趋势的统计量。
- 离群值 (Outliers): 分别输出 5 个极大值和极小值。
- 百分位数 (Percentiles): 输出变量 5%、10%、25%、50%、75%、90%、95% 分位数。

图 (Plots) 设置: 单击“图 (Plots)”按钮则弹出如图 5-17 所示的“图”对话框, 此对话框用于设置图形的各种特征。

- 箱图 (Boxplots): 定义箱图的输出。可以是按因子级别分组绘制 (Factor Levels together), 也可以不分组 (Dependent together), 或者不绘制 (None)。
- 描述图 (Descriptive): 定义是否输出茎叶图 (Stem-and-leaf) 和直方图 (Histogram)。
- 含检验的正态图 (Normality Plots with Tests): 选择是否进行正态检验, 且是否输出相应的 Q-Q 图。
- 含莱文检验的分布-水平图 (Spread vs Level with Levene Test): 当选入分组变量时, 该功能才被激活, 主要用于比较各组之间的离散程度是否一致。

选项 (Options) 设置: 单击“选项 (Options)”按钮, 弹出如图 5-18 所示的“选项设置”对话框, 此对话框用于设置缺失值处理方法。



图 5-17 “图 (Plots)”对话框

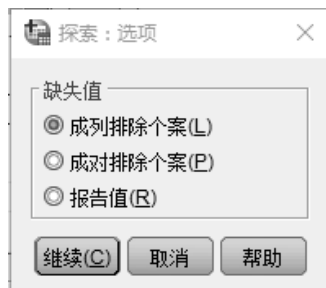




图 5-18 “选项 (Options) 设置”对话框

自助抽样 (Bootstrap) 设置: Bootstrap 方法可以导出稳健的标准误差估计值, 并能为诸如平均值、中位数、比例、概率比、相关系数或回归系数等估计值导出置信区间。它还可用于构建假设检验。

5.4.2 实例分析

本实例所用 SPSS 自带的数据集为 ceramics.sav, 数据集包含 5 个变量, 数据集格式如图 5-19 所示。

-  **结果文件** —— 附带光盘“PROGRAM\CH05\实例 5-3”文件夹
-  **动画演示** —— 附带光盘“AVI\实例 5-3.avi”文件

1. 参数设置

选择菜单“分析 (Analyze) 描述统计 (Descriptive Statistics) 探索 (Explore)”，弹出如图 5-20 所示的“探索变量设置”对话框，选择变量 Degrees Centigrade 到“因变量列表 (Dependent List)”选项栏中，选择变量 Alloy 到“因子列表 (Factor List)”选项栏中，选择变量 labrunid 到“个案标注依据 (Label Cases by)”选项栏中。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|----------|-----|----|------|-------------------|----------------|----|---|----|----|----|
| 1 | id | 数字 | 3 | 0 | Unit ID number | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 2 | lab | 数字 | 2 | 0 | Production Lab | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 3 | batch | 数字 | 2 | 0 | Alloy | {1, Premium... | 无 | 8 | 右 | 标度 | 输入 |
| 4 | temp | 数字 | 8 | 2 | Degrees Centig... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 5 | labrunid | 字符串 | 4 | 0 | | 无 | 无 | 8 | 左 | 标度 | 输入 |

图 5-19 ceramics.sav 的数据格式

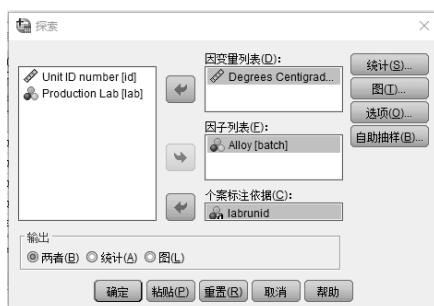


图 5-20 “探索 (Explore) 变量设置”对话框

然后单击“统计量 (Statistics)”按钮，弹出如图 5-21 所示的“统计量”对话框，选择“M-估计量 (M-estimators)”和“离群值 (Outliers)”选项栏，然后单击“继续”按钮返回主界面。

单击图 5-20 中的“图 (Plots)”按钮，弹出如图 5-22 所示的“图设置”对话框，选择“含检验的正态图 (Normality Plots with Tests)”选项，然后单击“继续”按钮返回主界面。

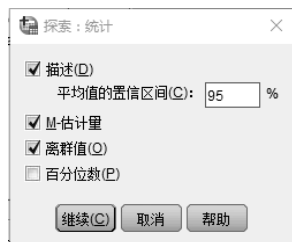


图 5-21 “统计量 (Statistics)”对话框



图 5-22 “图 (Plots) 设置”界面

2. 结果分析

设置好上述参数以后，单击主界面 Explore dialog box 中的“确定 (OK)”按钮进行分析，结果如下。首先是个案处理分析结果，如图 5-23 所示，包括观测量、缺失值等信息。

| 个案处理摘要 | | | | | | | |
|--------------------|----------|-----|--------|------|------|-----|--------|
| | | 有效 | | 个案缺失 | | 总计 | |
| Alloy | | 个案数 | 百分比 | 个案数 | 百分比 | 个案数 | 百分比 |
| Degrees Centigrade | Premium | 240 | 100.0% | 0 | 0.0% | 240 | 100.0% |
| | Standard | 240 | 100.0% | 0 | 0.0% | 240 | 100.0% |

图 5-23 个案处理分析结果

然后是描述性统计量，如图 5-24 所示，包含均值、95%置信区间、方差、中位数、标准差、最大最小值、偏度和峰度等信息。

| 描述 | | | | | |
|--------------------|----------|---------------|-----------|-----------|--|
| Alloy | | 统计 | | 标准误差 | |
| Degrees Centigrade | Premium | 平均值 | 1542.0787 | .61165 | |
| | | 平均值的 95% 置信区间 | 下限 | 1540.8738 | |
| | | | 上限 | 1543.2836 | |
| | | 5% 剪除后平均值 | 1541.2805 | | |
| | | 中位数 | 1539.7181 | | |
| | | 方差 | 89.789 | | |
| | | 标准差 | 9.47569 | | |
| | | 最小值 | 1530.44 | | |
| | | 最大值 | 1591.04 | | |
| | | 全距 | 60.61 | | |
| | | 四分位距 | 11.51 | | |
| | | 偏度 | 1.439 | .157 | |
| | | 峰度 | 3.036 | .313 | |
| | Standard | 平均值 | 1514.6564 | .62004 | |
| | | 平均值的 95% 置信区间 | 下限 | 1513.4350 | |
| | | | 上限 | 1515.8779 | |
| | | 5% 剪除后平均值 | 1514.7302 | | |
| | | 中位数 | 1514.5317 | | |
| | | 方差 | 92.269 | | |
| | | 标准差 | 9.60566 | | |
| | | 最小值 | 1488.30 | | |
| | | 最大值 | 1537.99 | | |
| | | 全距 | 49.69 | | |
| | | 四分位距 | 13.51 | | |
| | | 偏度 | -.078 | .157 | |
| | | 峰度 | -.343 | .313 | |

图 5-24 描述性统计量

然后是正态性检验结果，如图 5-25 所示，其中 Premium 变量对应的 K-S 检验 P 值和 Shapiro-Wilk 检验 P 值均为 0.000，所以，此变量的数据分布可以认为是正态的。而变量 Standard 数据的分布不能判断为正态分布。

| 正态性检验 | | | | | | | |
|--------------------|----------|--------------------------|-----|-------------------|---------|-----|------|
| | | 柯尔莫戈洛夫-斯米诺夫 ^a | | | 夏皮洛-威尔克 | | |
| Alloy | | 统计 | 自由度 | 显著性 | 统计 | 自由度 | 显著性 |
| Degrees Centigrade | Premium | .123 | 240 | .000 | .888 | 240 | .000 |
| | Standard | .027 | 240 | .200 [*] | .995 | 240 | .602 |

*. 这是真显著性的下限。
a. 里利氏显著性修正

图 5-25 正态性检验结果

然后是观测值的 Q-Q 图,如图 5-26 所示,从图中可以看出观测值的中间部分偏差并不大,而在观测图的两端均偏离直线。

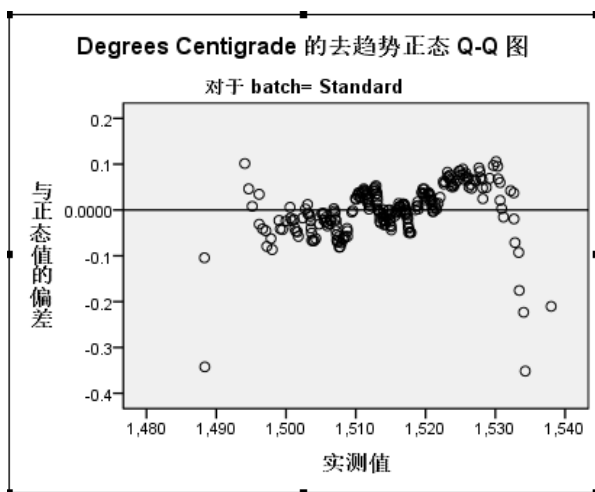


图 5-26 Q-Q 图

然后是箱图,如图 5-27 所示,其中变量 Premium 中有部分异常数据,数据偏大。

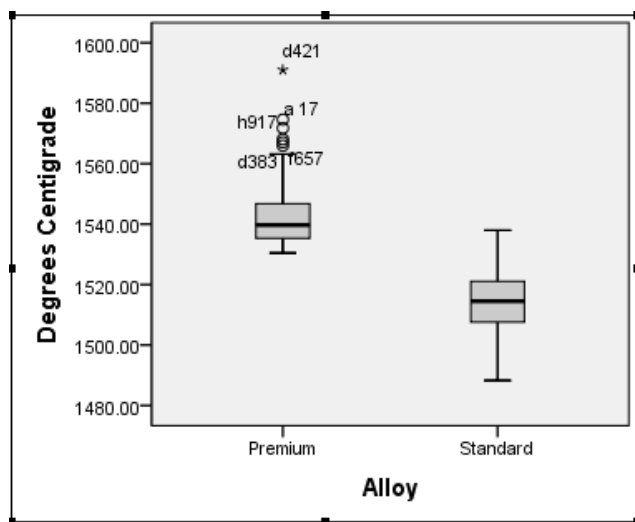


图 5-27 箱图

5.5 交叉表分析过程

交叉表给出了多个变量在不同取值下的数据分布,从而分析变量之前的相互关系。

5.5.1 交叉表过程的参数设置

选择菜单“分析 (Analyze) 描述统计 (Descriptive Statistics) 交叉表 (Crosstabs)”,弹出如图 5-28 所示的“交叉表”对话框,该对话框主要组成部分如下所述。



图 5-28 “交叉表 (Crosstabs)”对话框

行 (Row (s)) 选项栏：用于选择列变量。

列 (Column (s)) 选项栏：用于选择行变量。

层 (Layer)：用于选择分层变量，用上一张 (Previous) 和下一张 (Next) 按钮控制分层的层数。

显示簇状条形图 (Display Clustered Bar Charts)：用于选择输出分组条形图。

排除表 (Suppress Tables)：选择是否禁止输出交叉表。

精确 (Exact) 设置：单击图 5-28 中的“精确 (Exact)”按钮则弹出如图 5-29 所示的“精确设置”对话框，此对话框主要用于定义确切概率的计算。

- 仅渐进法 (Asymptotic only)：只计算近似概率。
- 蒙特卡洛法 (Monte Carlo)：用蒙特卡洛法计算精确概率，可以自行设置置信水平和样本数。
- 精确 (Exact)：在给定时间内计算精确概率。



图 5-29 “精确 (Exact) 设置”对话框

统计量 (Statistics) 设置：单击图 5-28 中的“统计量 (Statistics)”按钮，则弹出

如图 5-30 所示的“统计量”对话框，用于设置统计量。

- 卡方 (Chi-square)：卡方统计量。
- 相关性 (Correlations)：计算交叉表的 Pearson 相关系数和 Spearson 相关系数。
- 名义 (Nominal)：定义分类变量的相关性，共包括四个指标列联系数 (Contingency coefficient)，其值界于 0~1、Phi 和克莱姆 V (Phi and Cramer's V)，Phi 在四格表卡方检验中界于 1~1，在 R*C 表检验中界于 0~1；克莱姆 V 则界于 0~1、Lambda，在自变量预测中用于反映比例缩减误差，其值为 1 时表明自变量预测应变量好，为 0 时表明自变量预测应变量差、不确定性系数 (Uncertainty coefficient)，以熵为标准的比例缩减误差，其值接近 1 时表明后一变量的信息很大程度来自前一变量，其值接近 0 时表明后一变量的信息与前一变量无关。
- 有序 (Ordinal)：用于定义有序变量的相关性指标。
- 按区间标定 (Nominal by Interval)：用于分类变量的检验。
- Kappa：内部一致性系数。
- 风险 (Risk)：相对危险度。
- 麦克尼马尔 (McNemar)：进行麦克尼马尔检验。
- 柯克兰和曼特-亨塞尔统计 (Cochran's and Mantel-Haenszel)：进行独立性和齐次性检验。

单元格 (Cells) 设置：单击“单元格 (Cells)”按钮。弹出如图 5-31 所示的“单元格设置”对话框，用于定义交叉表中需要计算和输出的指标。

- 计数 (Counts)：定义输出频数。
- z-检验。
- 百分比 (Percentages)：定义需要计算的百分比。
- 残差 (Residuals)：定义输出的残差。
- 非整数权重 (Noninteger Weights)：当频数因为加权而变成小数的时候，选择该项对频数进行取整，主要包括 5 种取整方法。



图 5-30 “统计量 (Statistics)”对话框



图 5-31 “单元格 (Cells) 设置”对话框

格式 (Format) 设置：单击“格式 (Format)”按钮，弹出如图 5-32 所示的“格式设置”对话框，用于定义变量的排列方式。

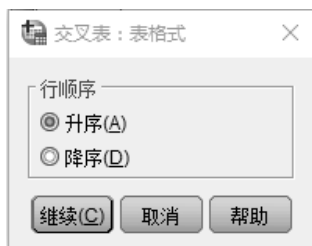


图 5-32 “格式 (Format) 设置”对话框

5.5.2 实例分析

本实例利用 SPSS 中自带的数据集 demo.sav，此数据集为某公司的报纸促销用户调查数据集，其数据集格式如图 5-33 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|----|---------|----|----|------|---------------------|------------------|----|---|----|----|----|
| 1 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 2 | marital | 数字 | 4 | 0 | Marital status | {0, Unmarrie... | 无 | 8 | 右 | 标度 | 输入 |
| 3 | address | 数字 | 4 | 0 | Years at curren... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 4 | income | 数字 | 8 | 2 | Household inco... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 5 | inccat | 数字 | 8 | 2 | Income categor... | {1.00, Under... | 无 | 8 | 右 | 有序 | 输入 |
| 6 | car | 数字 | 8 | 2 | Price of primary... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 7 | carcat | 数字 | 8 | 2 | Primary vehicle... | {1.00, Econ... | 无 | 8 | 右 | 有序 | 输入 |
| 8 | ed | 数字 | 4 | 0 | Level of education | {1, Did not c... | 无 | 8 | 右 | 标度 | 输入 |
| 9 | employ | 数字 | 4 | 0 | Years with curr... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 10 | retire | 数字 | 4 | 0 | Retired | {0, No}... | 无 | 8 | 右 | 标度 | 输入 |

图 5-33 数据集格式

- 结果文件** —— 附带光盘“PROGRAM\CH05\实例 5-4”文件夹
- 动画演示** —— 附带光盘“AVI\实例 5-4.avi”文件

1. 参数设置

选择菜单“分析 (Analyze) 描述统计 (Descriptive Statistics) 交叉表 (Crosstabs)”，弹出如图 5-34 所示的“交叉表选项栏设置”对话框，选择变量 Newspaper subscription 到“行 (Row(s))”选项栏中，选择变量 Response 到“列 (Column(s))”选项栏中。

然后单击“统计量 (Statistics)”按钮，则弹出如图 5-35 所示的“统计量设置”对话框，选择“风险 (Risk)”选项。然后单击“继续”按钮返回主界面。

2. 结果分析



图 5-35 “统计量 (Statistics) 设置”对话框

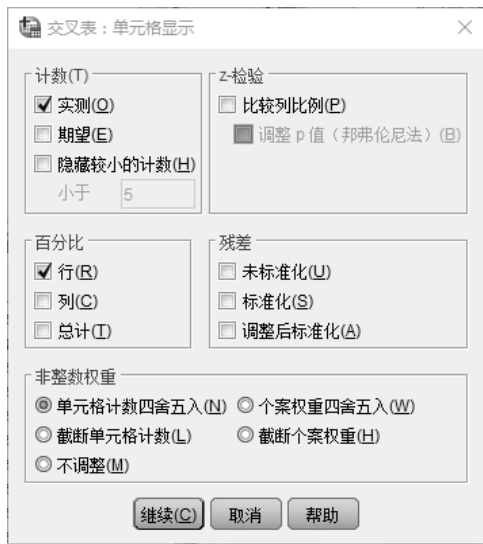


图 5-36 “单元格 (Cells) 设置”对话框

设置好上述参数以后，则单击“确定”按钮进行分析。首先是个案处理结果，如图 5-37 所示，包括观测量、缺失值等信息。

| 个案处理摘要 | | | | | | |
|--------------------------------------|------|--------|------|------|------|--------|
| | 有效 | | 个案缺失 | | 总计 | |
| | N | 百分比 | N | 百分比 | N | 百分比 |
| Newspaper subscription * Response | 6400 | 100.0% | 0 | 0.0% | 6400 | 100.0% |

图 5-37 个案处理结果

然后是交叉表，如图 5-38 所示，表中给出了反馈基本信息，Yes 信息中反馈的为 380，占 13.7%，没有订购对应的 No 信息中反馈的为 299，占 8.2%。总计反馈的数量为 679，占 10.6%。

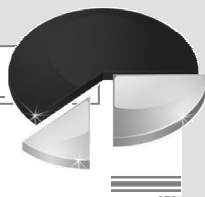
| Newspaper subscription * Response 交叉表 | | | | | |
|---------------------------------------|-------------------------------|-------------------------------|----------|-------|--------|
| | | | Response | | 总计 |
| | | | Yes | No | |
| Newspaper subscription | Yes | 计数 | 380 | 2388 | 2768 |
| | | 占 Newspaper subscription 的百分比 | 13.7% | 86.3% | 100.0% |
| | No | 计数 | 299 | 3333 | 3632 |
| | | 占 Newspaper subscription 的百分比 | 8.2% | 91.8% | 100.0% |
| 总计 | 计数 | | 679 | 5721 | 6400 |
| | 占 Newspaper subscription 的百分比 | | 10.6% | 89.4% | 100.0% |

图 5-38 交叉表

最后是风险估计值，如图 5-39 所示。此相对风险即是事件发生概率的比例，如 $13.7\%/8.2\%=1.668$ 。

| 风险评估 | | | |
|---|-------|----------|-------|
| | 值 | 95% 置信区间 | |
| | | 下限 | 上限 |
| Newspaper subscription (Yes / No) 的比值比 | 1.774 | 1.511 | 2.082 |
| 对于 cohort Response = Yes | 1.668 | 1.445 | 1.924 |
| 对于 cohort Response = No | .940 | .924 | .957 |
| 有效个案数 | 6400 | | |

图 5-39 风险估计



第 6 章 参 数 检 验

假设检验是先假设总体的分布形式或总体的参数具有某种特征，然后利用样本提供的信息来推断所提出的假设的正确性。这种处理问题的方法称为假设检验。

本章详细讲述利用 SPSS 软件进行参数检验的过程。



本讲内容

- 参数估计和假设检验概述
- 平均值 (Means) 过程
- 单样本 t 检验
- 独立两样本 t 检验
- 成对样本 t 检验

6.1 参数估计和假设检验概述

6.1.1 参数估计

在许多实际问题中，总体被理解为我们所研究的那个统计指标，它在一定范围内取数值，而且是以一定的概率取各种数值的，从而形成一个概率分布，但是这个概率分布往往是未知的。例如，为了制定绿色食品的有关规定，需要研究蔬菜中残留农药的分布状况，对这个分布我们知之甚少，以致它属于何种类型都不清楚。有时我们可以断定分布的类型，例如，在农民收入调查中，根据实际经验和理论分析如概率论中的中心极限定理，断定收入服从正态分布，但分布中的参数取何值却是未知的。这就导致统计估计问题。统计估计问题专门研究由样本估计总体的未知分布或分布中的未知参数。直接对总体的未知分布进行估计的问题称为非参数估计；当总体分布类型已知，仅需对分布的未知参数进行估计的问题称为参数估计。本节研究参数估计问题。本节及以后假设抽样方法为放回简单随机抽样，样本的每个分量都与总体同分布，它们之间相互独立。

1. 参数估计的基本方法

(1) 估计量与估计值

参数估计就是用样本统计量去估计总体参数。

用来估计总体参数的统计量的名称称为估计量，如样本平均值、样本比例、样本方差等都可以是一个估计量。

估计量的具体数值称为估计值。

(2) 点估计与区间估计

参数估计方法有点估计与区间估计两种方法。

1) 点估计法

设总体 X 的分布类型已知，但包含未知参数 θ ，从总体中抽取一个简单随机样本 (X_1, X_2, \dots, X_n) ，欲利用样本提供的信息对总体未知参数 θ 进行估计。构造一个适当的统计量

$$\hat{\theta} = T(X_1, X_2, \dots, X_n)$$

作为 θ 的估计，称 $\hat{\theta}$ 为未知参数 θ 的点估计量 (Point Estimate)。当有了一个具体的样本观察值 (x_1, x_2, \dots, x_n) 后，将其代入估计量中就得到估计量的一个具体观察值 $T(x_1, x_2, \dots, x_n)$ ，称为参数 θ 的一个点估计值。今后点估计量和点估计值这两个名词将不强调它们的区别，通称为点估计，根据上下文不不知道此处的点估计究竟是点估计量还是点估计值。

通俗地说，用样本估计量的值直接作为总体参数的估计值称为点估计。常用的点估计量为

$$\hat{\mu} = \bar{X} ; \hat{p} = P ; \hat{\sigma}^2 = s^2 = \frac{\sum (X - \bar{X})^2}{n-1}$$

2) 区间估计法

在参数估计中，虽然点估计可以给出未知参数的一个估计，但不能给出估计的精度。为此人们希望利用样本给出一个范围，要求它以足够大的概率包含待估参数真值。这就是导致区间估计 (Interval Estimation) 问题产生的原因。

区间估计，就是估计总体参数的区间范围，并要求给出区间估计成立的概率值。

设 θ 是未知参数， (X_1, X_2, \dots, X_n) 是来自总体的样本，构造两个统计量 $\hat{\theta}_1 = T_1(X_1, X_2, \dots, X_n)$ ， $\hat{\theta}_2 = T_2(X_1, X_2, \dots, X_n)$ ，对于给定的 α ($0 < \alpha < 1$)，若 $\hat{\theta}_1$ 、 $\hat{\theta}_2$ 满足

$$P\{\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2\} = 1 - \alpha$$

则称随机区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 是参数 θ 的置信水平 (Confidence Level) 为 $1 - \alpha$ 的置信区间 (Confidence Interval)， $1 - \alpha$ 称为 $[\hat{\theta}_1, \hat{\theta}_2]$ 的置信水平， $\hat{\theta}_1$ 、 $\hat{\theta}_2$ 称为置信限 (Confidence Limit)。

(3) 总体平均值的区间估计

1) 正态总体且方差已知；或非正态总体、方差未知、大样本情况下

在这种情况下，样本平均值的抽样分布呈正态分布，其数学期望为总体平均值 μ ，方差为 $\frac{\sigma^2}{n}$ 。则 $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ 称为总体平均值在 $1 - \alpha$ 置信水平下的置信区间。

设样本 (X_1, X_2, \dots, X_n) 来自正态总体 $N(\mu, \sigma_{\bar{X}}^2)$, μ 是总体平均值, 当 $\sigma_{\bar{X}}^2$ 已知时数理统计证明 \bar{X} 服从正态分布 $N(\mu, \frac{\sigma^2}{n})$, 从而 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ 服从标准正态分布 $N(0,1)$, 对给定的置信水平 $1-\alpha$ 查 $N(0,1)$ 表可得 $Z_{\frac{\alpha}{2}}$, 使得 $P\left\{\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq Z_{\frac{\alpha}{2}}\right\} = 1-\alpha$, 从而有 $P\left\{\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$, 取 $\hat{\mu}_1 = \bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \hat{\mu}_2 = \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$, 则 $[\hat{\mu}_1, \hat{\mu}_2]$ 即是 μ 的置信水平为 $1-\alpha$ 的置信区间。

2) 正态总体、方差未知、小样本情况下

如果总体服从正态分布, 无论样本容量大小, 样本平均值的抽样分布都服从正态分布。只要总体方差已知, 即使在小样本情况下, 也可以计算总体平均值的置信区间。如果总体方差 σ^2 未知, 需用样本方差 S^2 代替, 在小样本情况下, 应用 t 分布来建立总体平均值的置信区间。

t 分布是类似正态分布的一种对称分布, 它通常要比正态分布平坦和分散。随着自由度的增大, t 分布逐渐趋于正态分布。

正态总体、方差未知、小样本情况下, 总体平均值在 $1-\alpha$ 置信水平下的置信区间为

$$\bar{X} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \quad (\text{重复抽样条件下}); \quad \bar{X} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (\text{不重复抽样条件下})$$

式中: $t_{\frac{\alpha}{2}}(n-1)$ 为 t 分布临界值。

(4) 总体比例的区间估计

在大样本 (一般经验规则为 $np \geq 5$ 和 $n(1-p) \geq 5$) 条件下, 样本比例的抽样分布可用正态分布近似。在这种情况下, 数理统计已经证明如下结论:

置信水平为 $1-\alpha$ 的置信区间为

$$p \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} \quad (\text{重复抽样}); \quad p \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}} \quad (\text{不重复抽样})$$

(5) 总体方差的区间估计

数理统计证明, 对于容量为 n 的正态总体样本方差 S^2 , 若总体方差为 σ^2 , 则 $\frac{(n-1)S^2}{\sigma^2}$ 服从自由度为 $n-1$ 的 χ^2 分布。对给定的置信系数 $1-\alpha$, 查 χ^2 分布表得上 $\frac{\alpha}{2}$ 分位点 $\chi_{\frac{\alpha}{2}}^2(n-1)$ 和下 $1-\frac{\alpha}{2}$ 分位点 $\chi_{1-\frac{\alpha}{2}}^2(n-1)$, 使得 $P\left\{\chi_{1-\frac{\alpha}{2}}^2(n-1) \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\frac{\alpha}{2}}^2(n-1)\right\} = 1-\alpha$, 从而有 $P\left\{\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)}\right\} = 1-\alpha$, 取 $\hat{\sigma}_1^2 = \frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2(n-1)}$, $\hat{\sigma}_2^2 = \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)}$, 则

$[\hat{\sigma}_1^2, \hat{\sigma}_2^2]$ 即是 σ^2 的置信水平为 $1-\alpha$ 的置信区间。即

$$\frac{(n-1)s^2}{\chi_{\frac{\alpha}{2}}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)}$$

6.1.2 假设检验

假设检验是先假设总体的分布形式或总体的参数具有某种特征, 然后利用样本提供的信息来推断所提出的假设的正确性。这种处理问题的方法称为假设检验。

1. 基本概念

下面阐述假设检验的基本思想及所涉及的基本概念。

(1) 原假设和备择假设

一般地, 设统计模型为 $\{P_\theta; \theta \in \Theta\}$, 关于总体分布中的参数 θ 的推测, 即 $H: \theta \in \bar{\Theta} \subset \Theta$ 称为假设, 其中 $\bar{\Theta}$ 是参数空间 Θ 的非空真子集。如果 $\bar{\Theta}$ 仅有一个参数, 即 $\bar{\Theta} = \{\theta_0\}$, 则称 H 为简单假设, 否则称为复合假设。

在一个假设检验中, 常常涉及两个假设, 所要检验的假设称为原假设或零假设, 记为 H_0 。而与 H_0 不相容的假设称为备择假设或对立假设, 记为 H_1 。对参数统计模型 $\{P_\theta; \theta \in \Theta\}$ 而言, 原假设和备择假设这对矛盾的统一体, 即

$$H_0: \theta \in \Theta_0, H_1: \theta \in \Theta_1$$

称为假设检验问题。

(2) 拒绝域、接受域、检验统计量和检验函数

由于样本平均值 \bar{x} 是总体平均值 μ 的一致最小方差无偏估计, 这样样本平均值 \bar{x} 在一定程度上就反映了总体平均值 μ 的大小, 因此可考虑用样本平均值 \bar{x} 来做推断。当原假设 H_0 成立时, 样本平均值 \bar{x} 与总体平均值 μ_0 相差不应过大, 即偏差 $|\bar{x} - \mu_0|$ 不应过大, 若偏差 $|\bar{x} - \mu_0|$ 相当大, 就有理由怀疑 H_0 不成立而拒绝 H_0 。又因为当原假设 H_0 成立时, 统计量 $U = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ 服从标准正态分布 $N(0,1)$, 因此, 衡量 $|\bar{x} - \mu_0|$ 的大小就等价于衡量 $|U|$ 的大小。现在的问题是当 $|U|$ 大到什么程度才算过大, 才有理由拒绝 H_0 , 这就需要规定一个界限 c 。因此, 就获得了判断 H_0 是否成立的一个法则: 当 H_0 成立时, 若 $|U| \geq c$, 拒绝 H_0 , 否则接受 H_0 , 其中 c 是一个待定的常数, 称 c 为检验的临界值。

从上述解释可知, 检验一个假设就是根据某一法则在原假设和备择假设之间做出选择, 而基于样本 x_1, x_2, \dots, x_n 做出拒绝 H_0 或接受 H_0 所依赖的法则称为检验。

其中不等式 $|U| \geq c$ 实际上是将样本空间划分为两部分 W 和 W^c , 因此, 一个检验就等同于将样本空间分成两个互不相交的子集 W 和 W^c , 当 $(x_1, x_2, \dots, x_n) \in W$ 时就拒绝 H_0 , 认为备择假设 H_1 成立, 而当 $(x_1, x_2, \dots, x_n) \in W^c$ 就接受 H_0 , 认为 H_0 成立。称 W 为拒绝域, W^c 为接受域。这样检验和拒绝域之间就建立起一一对应关系。

(3) 两类错误、功效和功效函数

在假设检验中做出拒绝或接受原假设推断的依据是样本。由于样本的随机性和局限性, 进行检验时不可避免地会出现误判而犯错误, 这种可能犯的误差分为两类。

是当原假设 H_0 本来为真时, 样本观测值却落入拒绝域 W , 错误地拒绝了 H_0 , 这种误差通常称为第一类错误或“弃真”误差, 其概率为 $\alpha(\theta) = P_\theta\{x \in W\}$, $\theta \in \Theta_0$ 。

是当 H_0 本来不成立时, 样本观测值却落入接受域 W^c , 错误地接受了 H_0 , 从而犯了“取伪”误差, 称为第二类误差, 其概率为 $\beta(\theta) = P_\theta\{x \notin W\} = 1 - P_\theta\{x \in W\}$, $\theta \in \Theta_1$ 。

定义 1: 称 H_0 不成立时拒绝 H_0 的概率, 即 $\gamma(\theta) = P_\theta\{x \in W\} = 1 - \beta(\theta)$, $\theta \in \Theta_1$, 为一个检验的功效。而犯第一类错误的概率和功效可以看作函数

$$g(\theta) = P_\theta\{x \in W\} = E_\theta(\varphi(x)), \quad \theta \in \Theta$$

的不同取值, 这个函数称为功效函数。

(4) 检验水平

在检验一个假设时, 自然希望犯两类错误的概率都尽可能的小。但当样本容量 n 固定时, 要减少犯第一类错误的概率, 就会增大犯第二类错误的概率; 反之, 若要减少犯第二类错误的概率, 就会增大犯第一类错误的概率。即当样本容量固定时, 不可能同时减少犯两类错误的概率, 这是一对不可调和的矛盾。只有增大样本容量才能同时使 α, β 都变小, 这在很多实际问题中是不现实的。

Neyman-Pearson 检验原理就是控制犯第一类错误的概率在给定的范围内, 寻找检验使得犯第二类错误的概率尽可能的小, 就是使检验的功效尽可能的大。这样就是在给定一个较小的数 $\alpha(0 < \alpha < 1)$ 时, 一般取为 0.01、0.05、0.1 等, 在满足 $P_\theta\{x \in W\} = E_\theta(\varphi(x)) \leq \alpha$, $\theta \in \Theta_0$ 的检验函数类中, 寻找使功效 $E_\theta(\varphi(x)) = P_\theta\{x \in W\}$, $\theta \in \Theta_1$ 尽可能大的检验函数。

定义 2: 对给定的 $\alpha \in (0, 1)$, 若检验函数 $\varphi(x)$ 对所有的参数 $\theta \in \Theta_0$, 满足 $E_\theta(\varphi(x)) \leq \alpha$, 则称 $\varphi(x)$ 是一个显著性水平为 α 的检验。

2. 正态总体的检验

(1) 单个正态总体方差已知时总体平均值的检验

设 x_1, x_2, \dots, x_n 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, 其中 σ^2 已知, 构造检验统计量, 即

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

当 $\mu = \mu_0$ 时, z 服从 $N(0, 1)$ 。给定显著性水平 α , 则有以下规则。

1) $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$

检验规则为当 $|z| \geq z_{\frac{\alpha}{2}}$ 时, 拒绝 H_0 ; 当 $|z| < z_{\frac{\alpha}{2}}$ 时, 不能拒绝 H_0 。

2) $H_0: \mu = \mu_0$, $H_1: \mu > \mu_0$

检验规则为当 $z \geq z_\alpha$ 时, 拒绝 H_0 ; 当 $z < z_\alpha$ 时, 不能拒绝 H_0 。

3) $H_0: \mu = \mu_0$, $H_1: \mu < \mu_0$

检验规则为当 $z \leq -z_\alpha$ 时, 拒绝 H_0 ; 当 $z > -z_\alpha$ 时, 不能拒绝 H_0 。

(2) 单个正态总体方差未知时总体平均值的检验

由抽样分布定理可知

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

故可以作为检验统计量，给定显著性水平 α ，则有以下规则。

1) $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$

检验规则为当 $|t| \geq t_{\frac{\alpha}{2}}(n-1)$ 时，拒绝 H_0 ；当 $|t| < t_{\frac{\alpha}{2}}(n-1)$ 时，不能拒绝 H_0 。

2) $H_0: \mu = \mu_0$, $H_1: \mu > \mu_0$

检验规则为当 $t \geq t_{\alpha}(n-1)$ 时，拒绝 H_0 ；当 $t < t_{\alpha}(n-1)$ 时，不能拒绝 H_0 。

3) $H_0: \mu = \mu_0$, $H_1: \mu < \mu_0$

检验规则为当 $t \leq -t_{\alpha}(n-1)$ 时，拒绝 H_0 ；当 $t > -t_{\alpha}(n-1)$ 时，不能拒绝 H_0 。

(3) 单个正态总体方差的检验

总体方差 σ^2 是用样本方差 s^2 来估计的。根据抽样分布理论，检验统计量，即

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

对给定的显著性水平 α ，则有以下规则。

1) $H_0: \sigma^2 = \sigma_0^2$, $H_1: \sigma^2 \neq \sigma_0^2$

检验规则为当 $\chi^2 \geq \chi_{\frac{\alpha}{2}}^2(n-1)$ 或 $\chi^2 \leq \chi_{1-\frac{\alpha}{2}}^2(n-1)$ 时拒绝 H_0 ，否则不能拒绝 H_0 。

2) $H_0: \sigma^2 = \sigma_0^2$, $H_1: \sigma^2 > \sigma_0^2$

检验规则为当 $\chi^2 \geq \chi_{\alpha}^2(n-1)$ 时拒绝 H_0 ，否则不能拒绝 H_0 。

3) $H_0: \sigma^2 = \sigma_0^2$, $H_1: \sigma^2 < \sigma_0^2$

检验规则为当 $\chi^2 \leq \chi_{1-\alpha}^2(n-1)$ 时拒绝 H_0 ，否则不能拒绝 H_0 。

这种用服从 χ^2 分布的统计量作为检验统计量的检验称为 χ^2 检验。

(4) 两个正态总体平均值的检验

设总体 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, x_1, x_2, \dots, x_{n_1} 和 y_1, y_2, \dots, y_{n_2} 分别是来自总体 X 和 Y 的样本，且两样本相互独立。样本平均值和样本方差分别记为 \bar{x}, S_1^2 和 \bar{y}, S_2^2 。考虑假设检验问题

$$H_0: \mu_1 = \mu_2, \quad H_1: \mu_1 \neq \mu_2$$

这个等价于假设检验问题

$$H_0: \mu_1 - \mu_2 = 0, \quad H_1: \mu_1 - \mu_2 \neq 0$$

这与单个总体情形十分类似，只要由样本构造出 $\mu_1 - \mu_2$ 的良好估计来进行比较就行了。下面分四种情形来讨论假设检验问题的显著性检验。

1) 方差 σ_1^2, σ_2^2 已知

由于 \bar{x} 和 \bar{y} 分别是 μ_1 和 μ_2 的一致最小方差无偏估计，因此 $\bar{x} - \bar{y}$ 是 $\mu_1 - \mu_2$ 的良好估计。当 H_0 成立时，差的绝对值 $|(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)|$ 不能太大，即 $|\bar{x} - \bar{y}|$ 不能太大，太大时就有理由怀疑 H_0 的正确性。从而拒绝域的形式为 $W = \{|\bar{x} - \bar{y}| > c\}$ ，其中 c 是待定的常数。

由于

$$\bar{x} - \bar{y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

当 H_0 成立时, 有

$$U = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

这样犯第一类错误的概率为

$$P_0\{W\} = P_0\{|\bar{x} - \bar{y}| > c\} = P_0\left\{|U| > \frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right\} = \alpha$$

由标准正态分布分位点的定义, 有

$$\frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = z_{1-\frac{\alpha}{2}}$$

因此, 可选取 U 作为检验统计量, 在显著性水平 α 下的拒绝域可简记为 $W = \{|U| > z_{1-\frac{\alpha}{2}}\}$,

这是个 u 检验。

2) 方差 σ_1^2, σ_2^2 未知, 但有 $\sigma_1^2 = \sigma_2^2 = \sigma^2$

构造检验统计量

$$t = \frac{\bar{x} - \bar{y}}{S_W \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

与 1) 推导类似, 可得显著性水平 α 下的拒绝域为

$$W = \{|t| > t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)\}$$

这是个 t 检验。

3) 方差 σ_1^2, σ_2^2 未知, $\sigma_1^2 \neq \sigma_2^2$, 但 $n_1 = n_2 = n$

这时令

$$z_i = x_i - y_i, \quad i = 1, 2, \dots, n$$

则有

$$E(z_i) = E(x_i - y_i) = \mu_1 - \mu_2 = \mu; \quad \text{Var}(z_i) = \text{Var}(x_i) + \text{Var}(y_i) = \sigma_1^2 + \sigma_2^2 = \sigma^2$$

这样可认为 z_1, z_2, \dots, z_n 是来自总体 Z 服从正态分布 $N(\mu, \sigma^2)$ 的样本。此时假设检验问题等价于假设检验问题

$$H_0: \mu = 0, \quad H_1: \mu \neq 0$$

这是单个正态总体方差 σ^2 未知下, 关于平均值 μ 的假设检验的特殊情形, 应该使用 t 检验, 即检验统计量为 $t = \frac{\bar{z}}{S/\sqrt{n}}$, 其中 \bar{z} 和 S^2 分别是样本 z_1, z_2, \dots, z_n 的样本平均值和样本方差。

在显著性水平 α 下, 假设检验问题的拒绝域为 $W = \left\{ |t| > t_{1-\frac{\alpha}{2}}(n-1) \right\}$, 这就是配对试验的 t 检验。

4) 方差 σ_1^2, σ_2^2 未知, $\sigma_1^2 \neq \sigma_2^2$, $n_1 \neq n_2$

下面介绍 Scheffe 方法。不妨假设 $n_1 < n_2$, 令

$$z_i = x_i - \sqrt{\frac{n_1}{n_2}} y_i + \frac{1}{\sqrt{n_1 n_2}} \sum_{j=1}^{n_1} y_j - \frac{1}{n_2} \sum_{j=1}^{n_2} y_j, \quad i=1, 2, \dots, n$$

则有

$$E(z_i) = \mu_1 - \mu_2 = \mu; \quad \text{Var}(z_i) = \sigma_1^2 + \frac{n_1}{n_2} \sigma_2^2; \quad \text{Cov}(z_i, z_j) = 0, \quad i \neq j, \quad i, j=1, 2, \dots, n_1$$

因此, 可认为 z_1, z_2, \dots, z_n 来自总体 Z 服从正态分布 $N\left(\mu, \sigma_1^2 + \frac{n_1}{n_2} \sigma_2^2\right)$ 的样本。此时, 关于两个总体平均值是否相等的假设检验问题就等价于假设检验问题, 即

$$H_0: \mu = 0, \quad H_1: \mu \neq 0$$

这仍是单个正态总体方差 $\sigma_1^2 + \frac{n_1}{n_2} \sigma_2^2$ 未知下, 关于平均值 μ 的假设检验的特殊情形, 应该

使用 t 检验, 即检验统计量为 $t = \frac{\bar{z}}{S/\sqrt{n_1}}$, 其中 \bar{z} 和 S^2 分别是样本 z_1, z_2, \dots, z_{n_1} 的样本平均值

和样本方差。在显著性水平 α 下, 假设检验问题的拒绝域为 $W = \left\{ |t| > t_{1-\frac{\alpha}{2}}(n_1-1) \right\}$ 。

(5) 两个正态总体方差之比的检验

设总体 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 假设检验问题的检验统计量为

$$F = \frac{S_1^2}{S_2^2}$$

服从 $F(n_1-1, n_2-1)$, 给定显著性水平 α , 则有以下规则。

1) $H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$

检验规则为当 $F \geq F_{\frac{\alpha}{2}}(n_1-1, n_2-1)$ 或 $F \leq F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1) = 1/F_{\frac{\alpha}{2}}(n_2-1, n_1-1)$ 时拒绝

H_0 , 否则不能拒绝 H_0 。

2) $H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 > \sigma_2^2$

检验规则为当 $F \geq F_{\alpha}(n_1-1, n_2-1)$ 时拒绝 H_0 , 否则不能拒绝 H_0 。

3) $H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 < \sigma_2^2$

检验规则为当 $F \leq 1/F_{\alpha}(n_2-1, n_1-1)$ 时拒绝 H_0 , 否则不能拒绝 H_0 。

6.2 平均值 (Means) 过程

Means 过程主要用于进行分组计算、比较变量的描述性统计量,还可以给出方差分析表和检验结果等信息。

6.2.1 SPSS 的平均值过程参数的设置

选择菜单“分析 (Analyze) 比较平均值 (Compare Means) 平均值 (Means)”,则弹出如图 6-1 所示的“平均值过程参数设置”对话框,其各项具体功能如下。

1. 变量设置

图 6-1 左边是待分析的变量列表。

- 因变量列表 (Dependent List): 选入待分析变量。
- 层 (Layer 1 of 1): 用于定义分组变量。
- 自变量列表 (Independent List): 选择分组变量。可以定义多层分组变量,每层分组变量中也可以有多个变量。
- 上一张 (Previous): 选择前一层的分组变量。
- 下一张 (Next): 选择下一层的分组变量。

2. 选项 (Options) 选项设置

单击图 6-1 所示的“选项 (Options)”按钮,则弹出如图 6-2 所示的“选项设置”对话框。



图 6-1 “平均值 (Means) 过程参数设置”对话框



图 6-2 “选项 (Options) 设置”对话框

统计量 (Statistics): 列出可以选择的描述性统计量。

单元格统计量 (Cell Statistics): 选择要输出的统计量,默认为平均值、个案数、标准差。

第一层的统计量 (Statistics for First Layer): 定义是否进行分组第一层变量的方差分



析 (Anova Table and Eta) 和线性相关度的检验 (Test for Linearity)。

6.2.2 平均值过程实例

本实例中所用数据集为 SPSS 自带的数据集 hourlywagedata.sav, 数据集是关于职位 (position)、年龄范围 (agerange)、时薪 (hourwage), 以及工作经验 (yrsscale) 变量的调查数据。数据集 hourlywagedata.sav 的格式如图 6-3 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|----------|----|----|------|------------------|-------------------|----|---|----|----|----|
| 1 | position | 数字 | 1 | 0 | Nurse Type | {0, Hospital}... | 无 | 8 | 右 | 标度 | 输入 |
| 2 | agerange | 数字 | 1 | 0 | Age Range | {1, 18-30}... | 无 | 8 | 右 | 标度 | 输入 |
| 3 | yrsscale | 数字 | 1 | 0 | Years Experience | {1, 5 or less}... | 无 | 8 | 右 | 标度 | 输入 |
| 4 | hourwage | 数字 | 6 | 2 | Hourly Salary | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 6-3 数据集 hourlywagedata.sav 的格式

-  **结果文件** —— 附带光盘 “PROGRAM\CH06\实例 6-1” 文件夹
-  **动画演示** —— 附带光盘 “AVI\实例 6-1.avi” 文件

1. 参数设置

选择菜单 “分析 (Analyze) 比较平均值 (Compare Means) 平均值 (Means)”, 则弹出如图 6-4 所示的 “平均值设置” 对话框, 其各项具体功能如下。选择变量 Hourly Salary 到 “因变量列表 (Dependent List)” 选项栏中, 选中变量 Years Experience 到 “自变量列表 (Independent List)” 选项栏中。

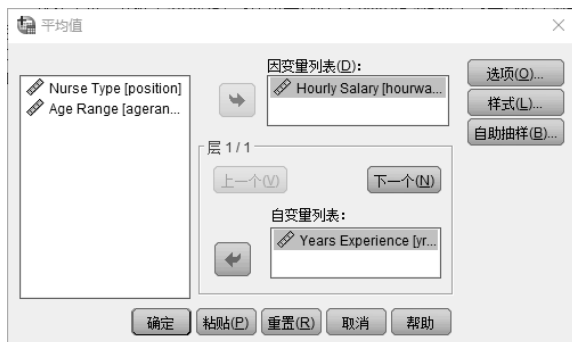


图 6-4 “平均值 (Means) 设置” 对话框

2. 结果分析

单击 “确定” 按钮进行分析, 结果如下。图 6-5 是基本的统计信息。样本包含个数为 2911, 已排除样本个数为 89。

| 个案处理摘要 | | | | | | |
|----------------------------------|------|-------|------|------|------|--------|
| | 包括 | | 个案排除 | | 总计 | |
| | 个案数 | 百分比 | 个案数 | 百分比 | 个案数 | 百分比 |
| Hourly Salary * Years Experience | 2911 | 97.0% | 89 | 3.0% | 3000 | 100.0% |

图 6-5 基本的统计信息

图 6-6 是统计分析详细的报告。分别给出了各组中的平均值、个数及标准差。

| 报告 | | | |
|------------------|---------|------|---------|
| Hourly Salary | | | |
| Years Experience | 平均值 | 个案数 | 标准差 |
| 5 or less | 18.0416 | 221 | 3.86667 |
| 6-10 | 18.9169 | 460 | 3.77816 |
| 11-15 | 19.6616 | 752 | 3.90528 |
| 16-20 | 20.2876 | 729 | 3.82786 |
| 21-35 | 21.2594 | 539 | 4.08669 |
| 36 or more | 21.6342 | 210 | 3.61826 |
| 总计 | 20.0159 | 2911 | 4.00309 |

图 6-6 统计分析详细的报告

6.3 单样本 t 检验

6.3.1 单样本 t 检验过程的参数设置

选择菜单“分析 (Analyze) 比较平均值 (Compare Means) 单样本 t 检验 (One-Sample t Test)”，则弹出如图 6-7 所示的“单样本 t 检验过程的参数设置”对话框，其各项具体功能如下所述。

1. 变量选择

进行 t 检验之前，则需要进行变量选择设置。图 6-7 中左边的待分析的变量列表。

- 检验变量 (Test Variable(s))：选入进行 t 检验的变量，可以选入多个变量。
- 检验值 (Test Value)：输入总体平均值。

图 6-7 “单样本 t 检验 (One-Sample t Test) 过程的参数设置”对话框

2. 选项 (Options) 设置

单击如图 6-7 所示的“选项 (Options)”按钮,则弹出如图 6-8 所示的“选项设置”对话框。

置信区间百分比 (Confidence Interval): 设置样本平均值与总体平均值之差的置信区间。

缺失值 (Missing Values): 缺失值处理方式。

- 按具体分析排除个案 (Exclude cases analysis by analysis): 仅当数据要分析的变量值缺失时才会剔除该数据。
- 成列排除个案 (Exclude cases listwise): 只要数据中有变量值缺失就剔除该数据。

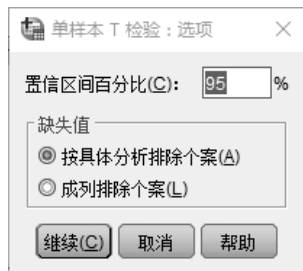


图 6-8 “选项设置”对话框

6.3.2 实例分析

本实例中所用数据集为 SPSS 自带的数据集 brakes.sav, 数据集是关于机器 (machine), 以及制动器 (brake) 变量的调查数据。数据集 brakes.sav 的格式如图 6-9 所示。

| 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 排序 | 存储 | 角色 |
|-----------|----|----|------|--------------------------|---|----|---|----|----|----|
| 1 machine | 数字 | 2 | 0 | Machine Number | 无 | 无 | 8 | 向右 | 存储 | 输入 |
| 2 brake | 数字 | 8 | 4 | Disc Brake Diameter (mm) | 无 | 无 | 8 | 向右 | 存储 | 输入 |

图 6-9 数据集 hourlywagedata.sav 的格式



结果文件

——附带光盘“PROGRAM\CH06\实例 6-2”文件夹



动画演示

——附带光盘“AVI\实例 6-2.avi”文件

1. 参数设置

首先对数据集进行预处理,选择菜单“数据 拆分文件 (Split File)”,弹出如图 6-10 所示的“分割文件设置”对话框。首先选择“比较组 (Compare Groups)”选项,再选择变量 Machine Number 到“分组方式 (Compare Based on) 变量框”中,然后单击“确定”按钮。

然后选择菜单“分析 (Analyze) 比较平均值 (Compare Means) 单样本 t 检验 (One-Sample t Test)”,则弹出如图 6-11 所示的“单样本 t 检验设置”对话框。选择变量 Disc Brake Diameter 到“检验变量 (Test Variable(s))”选项栏中,在下面的检验值 (Test Value) 选项栏中填入 322。

单击图 6-11 中的“选项 (Options)”按钮,弹出如图 6-12 所示的“选项设置”对话框,在置信区间 (Confidence Interval) 选项栏中填入 90%。然后单击“继续”按钮返回主界面。

2. 结果分析

单击主界面单样本 t 检验 (One-Sample t Test) 的“确定”按钮进行 t 检验分析。结果

如下。图 6-13 给出的是基本的样本统计量，包括平均值、标准差、标准误等信息。



图 6-10 “分割文件 (Split File) 设置”对话框

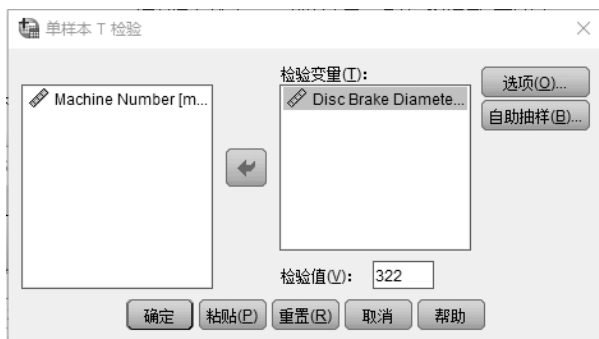


图 6-11 “单样本 t 检验设置”对话框

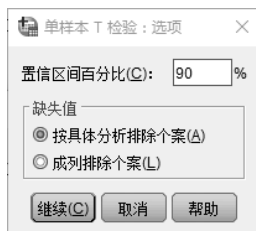


图 6-12 “选项 (Option) 设置”对话框

| 单样本统计 | | | | | |
|----------------|--------------------------|-----|------------|----------|----------|
| Machine Number | | 个案数 | 平均值 | 标准差 | 标准误差平均值 |
| 1 | Disc Brake Diameter (mm) | 16 | 321.998514 | .0111568 | .0027892 |
| 2 | Disc Brake Diameter (mm) | 16 | 322.014263 | .0106913 | .0026728 |
| 3 | Disc Brake Diameter (mm) | 16 | 321.998283 | .0104812 | .0026203 |
| 4 | Disc Brake Diameter (mm) | 16 | 321.995435 | .0069883 | .0017471 |
| 5 | Disc Brake Diameter (mm) | 16 | 322.004249 | .0092022 | .0023005 |
| 6 | Disc Brake Diameter (mm) | 16 | 322.002452 | .0086440 | .0021610 |
| 7 | Disc Brake Diameter (mm) | 16 | 322.006181 | .0093303 | .0023326 |
| 8 | Disc Brake Diameter (mm) | 16 | 321.996699 | .0077085 | .0019271 |

图 6-13 基本的样本统计量

图 6-14 输出的是 t 检验结果，包括检验的总体平均值、 t 统计量等信息。第 1 组、第 3 组、第 5 组、第 6 组、第 8 组中的检验结果显著性值分别为 0.602、0.522、0.085、0.274、0.107 均大于 0.05，所以在显著性水平 0.05 下，变量 Disc Brake Diameter (mm) 与平均值 322 有显著差异。第 2 组、第 4 组、第 7 组的变量 Disc Brake Diameter (mm) 与平均值 322 并无显著差异。

| 单样本检验 | | | | | | |
|----------------|--------------------------|--------|-----|----------|-----------|----------------------|
| Machine Number | | t | 自由度 | 显著性 (双尾) | 平均值差值 | 差值 90% 置信区间 下限 上限 |
| 1 | Disc Brake Diameter (mm) | -.533 | 15 | .602 | -.0014858 | -.006375 .003404 |
| 2 | Disc Brake Diameter (mm) | 5.336 | 15 | .000 | .0142629 | .009577 .018948 |
| 3 | Disc Brake Diameter (mm) | -.655 | 15 | .522 | -.0017174 | -.006311 .002876 |
| 4 | Disc Brake Diameter (mm) | -2.613 | 15 | .020 | -.0045649 | -.007628 -.001502 |
| 5 | Disc Brake Diameter (mm) | 1.847 | 15 | .085 | .0042486 | .000216 .008282 |
| 6 | Disc Brake Diameter (mm) | 1.134 | 15 | .274 | .0024516 | -.001337 .006240 |
| 7 | Disc Brake Diameter (mm) | 2.650 | 15 | .018 | .0061813 | .002092 .010270 |
| 8 | Disc Brake Diameter (mm) | -1.713 | 15 | .107 | -.0033014 | -.006680 .000077 |

图 6-14 t 检验结果

6.4 独立样本 t 检验

6.4.1 独立样本 t 检验过程的参数设置

选择菜单“分析 (Analyze) 比较平均值 (Compare Means) 独立样本 t 检验 (Independent-Sample T Test)”，则弹出如图 6-15 所示的“独立样本 t 检验”过程的参数设置对话框，其各项具体功能如下所述。

图 6-15 “独立样本 t 检验”过程的参数设置对话框

1. 变量选项

图 6-15 中左边为候选变量列表框，该变量框只显示可以进行 t 检验的变量。

- 检验变量 (Test Variable(s))：选入进行 t 检验的变量。
- 分组变量 (Grouping Variables)：选入分组变量，选入变量后，则激活其后的“定义组 (Define Groups)”按钮，单击后则弹出如图 6-16 所示的“定义组”设置对话框。使用指定值 (Use specified values) 用于特定的变量值分组，当变量的取值等于组 1 (Group1) 框中自定义值时将其划分为第一组；取值等于组 2 (Group2) 框中自定义值时将其划分为第二组。割点 (Cut Point) 用于定义分割点值。

2. 选项 (Options) 设置

单击如图 6-15 所示的“选项 (Options)”按钮,则弹出如图 6-17 所示的“独立样本 T 检验:选项”设置对话框。

置信区间百分比 (Confidence Interval): 设置样本平均值与总体平均值之差的置信区间,系统默认为 95%。

缺失值 (Missing Values): 缺失值处理方式。

- 按具体分析排除个案 (Exclude cases analysis by analysis): 仅当数据要分析的变量值缺失时才会剔除该数据。
- 成列排除个案 (Exclude cases listwise): 数据中有变量值缺失时剔除该数据。

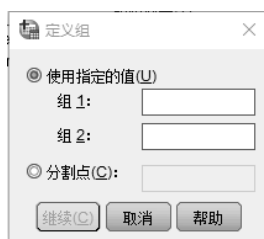


图 6-16 “定义组”设置对话框

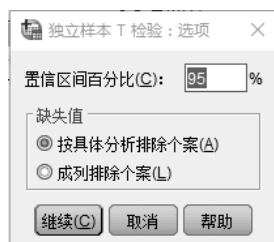


图 6-17 “独立样本 T 检验:选项”设置对话框

6.4.2 实例分析

本实例中所用数据集为 SPSS 自带的数据集 creditpromo.sav, 数据集有 3 个变量, 500 个观测样本。数据集 creditpromo.sav 的格式如图 6-18 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|---------|----|----|------|---------------------|---|----|---|----|----|----|
| 1 | id | 数字 | 12 | 0 | Customer ID | 无 | 无 | 8 | 右 | 名义 | 输入 |
| 2 | insert | 数字 | 1 | 0 | Type of mail ins... | 无 | 无 | 8 | 右 | 名义 | 输入 |
| 3 | dollars | 数字 | 8 | 2 | \$ spent during ... | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 6-18 数据集 creditpromo.sav 的格式



结果文件

——附带光盘“PROGRAM\CH06\实例 6-3”文件夹



动画演示

——附带光盘“AVI\实例 6-3.avi”文件

1. 参数设置

选择菜单“分析 (Analyze) 比较平均值 (Compare Means) 独立样本 t 检验 (Independent-Sample t Test)”, 则弹出如图 6-19 所示的“独立两样本 t 检验”对话框。选择变量 \$ spent during promotional period 到“检验变量 (Test Variables)”变量框中。同时选择变量

Type of mail insert received 到“分组变量 (Grouping Variables)”变量框中。



图 6-19 “独立两样本 t 检验”对话框

然后单击“定义组 (Define Groups)”按钮，弹出如图 6-20 所示的“定义组”对话框，在选项栏组 1 (Group 1) 和组 2 (Group 2) 中分别填写 0 和 1，然后单击“继续 (continue)”按钮返回主界面。

2. 结果分析

设置好参数以后，则单击主界面独立样本 t 检验中的“确定 (OK)”按钮进行分析。结果如下。如图 6-21 所示为基本统计量，包括观察样本数、平均值、标准差和标准误。

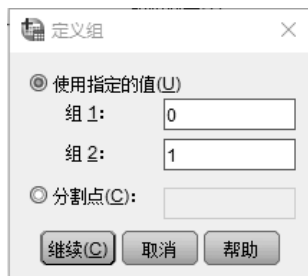


图 6-20 “定义组”对话框

| 组统计 | | | | | |
|------------------------------------|------------------------------|-----|-----------|-----------|----------|
| | Type of mail insert received | 个案数 | 平均值 | 标准差 | 标准误差平均值 |
| \$ spent during promotional period | Standard | 250 | 1566.3890 | 346.67305 | 21.92553 |
| | New Promotion | 250 | 1637.5000 | 356.70317 | 22.55989 |

图 6-21 基本统计量

图 6-22 输出的是独立样本检验结果，给出了关于方差齐性的 Levene 检验结果和关于平均值相等的 t 检验结果。 F 统计量的显著性值为 0.276 大于 0.10，所以不可否认方差相等的假设，第一行 t 检验的显著性值为 0.024 小于 0.10，所以在显著性水平 0.10 下，认为通过 Email 促销可以提高消费额。

| 独立样本检验 | | | | | | | | | |
|------------------------------------|--------|-------|------|---------------|---------|----------|-----------|----------|----------------------|
| 莱文方差等同性检验 | | | | 平均值等同性 t 检验 | | | | | |
| | | F | 显著性 | t | 自由度 | 显著性 (双尾) | 平均值差值 | 标准误差值 | 差值 95% 置信区间 下限 上限 |
| \$ spent during promotional period | 假定等方差 | 1.190 | .276 | -2.260 | 498 | .024 | -71.11095 | 31.45914 | -132.91995 -9.30196 |
| | 不假定等方差 | | | -2.260 | 497.595 | .024 | -71.11095 | 31.45914 | -132.92007 -9.30183 |

图 6-22 独立样本检验结果

6.5 成对样本 t 检验

6.5.1 成对样本 t 检验过程的参数设置

选择菜单“分析 (Analyze) 比较平均值 (Compare Means) 成对样本 t 检验 (Paired-Sample t Test)”, 则弹出如图 6-23 所示的“成对样本 t 检验”过程的参数设置对话框, 其各项具体功能如下所述。

1. 变量选项

图 6-23 中左边为候选变量列表框, 该变量框只显示可以进行 t 检验的变量。

- 配对变量 (Paired Variables): 选入进行 t 检验的配对样本。



图 6-23 “成对样本 t 检验”过程的参数设置对话框

2. 选项 (Options) 设置

单击如图 6-23 所示的“选项 (Options)”按钮, 则弹出如图 6-24 所示的“选项”设置对话框。

置信区间百分比 (Confidence Interval): 设置样本平均值与总体平均值之差的置信区间, 系统默认为 95%。

缺失值 (Missing Values): 缺失值处理方式。

- 按具体分析排除个案 (Exclude cases analysis by analysis): 仅当数据要分析的变量值缺失时才会剔除该数据。
- 成列排除个案 (Exclude cases listwise): 只要数据中有变量值缺失就剔除该数据。

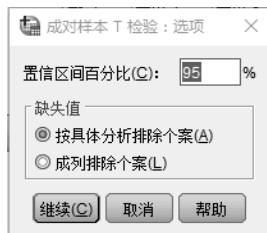




图 6-24 “选项”设置对话框

6.5.2 实例分析

本实例中所用数据集为 SPSS 自带的数据集 dietstudy.sav, 数据集有 13 个变量, 16 个观测。数据集 dietstudy.sav 的格式如图 6-25 所示。

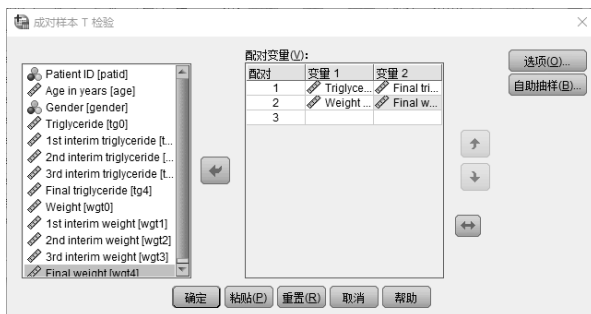
| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|--------|----|----|------|----------------------|--------------|----|---|----|----|----|
| 1 | patid | 数字 | 4 | 0 | Patient ID | 无 | 无 | 8 | 右 | 名义 | 输入 |
| 2 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 3 | gender | 数字 | 4 | 0 | Gender | {0, Male}... | 无 | 8 | 右 | 名义 | 输入 |
| 4 | tg0 | 数字 | 4 | 0 | Triglyceride | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 5 | tg1 | 数字 | 4 | 0 | 1st interim trigl... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 6 | tg2 | 数字 | 4 | 0 | 2nd interim trigl... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 7 | tg3 | 数字 | 4 | 0 | 3rd interim trigl... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 8 | tg4 | 数字 | 4 | 0 | Final triglyceride | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 6-25 数据集 dietstudy.sav 的格式

-  **结果文件** —— 附带光盘 “PROGRAM\CH06\实例 6-4” 文件夹
-  **动画演示** —— 附带光盘 “AVI\实例 6-4.avi” 文件

1. 参数设置

选择菜单“分析 (Analyze) 比较平均值 (Compare Means) 成对样本 t 检验 (Paired-Sample t Test)”，则弹出如图 6-26 所示的“成对样本 t 检验”设置对话框，其各项具体功能如下。选择变量 Triglyceride 和 Final Triglyceride 到“成对变量 (Paired Variables)”选项框中。同时选择变量 Weight 和 Final Weight 到“成对变量 (Paired Variables)”选项框中。

图 6-26 “成对样本 t 检验”设置对话框

2. 结果分析

设置好各个参数以后，则单击“确定 (OK)”按钮进行分析。首先是基本的统计分析，如图 6-27 所示。包括平均值、观测变量、标准差和标准误。

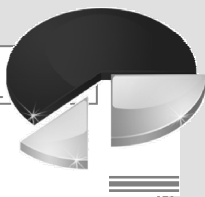
| 配对样本统计 | | | | | |
|--------|--------------------|--------|-----|--------|---------|
| | | 平均值 | 个案数 | 标准差 | 标准误差平均值 |
| 配对 1 | Triglyceride | 138.44 | 16 | 29.040 | 7.260 |
| | Final triglyceride | 124.38 | 16 | 29.412 | 7.353 |
| 配对 2 | Weight | 198.38 | 16 | 33.472 | 8.368 |
| | Final weight | 190.31 | 16 | 33.508 | 8.377 |

图 6-27 基本的统计分析

图 6-28 所示的是独立样本 t 检验，给出了平均值的检验结果，有置信区间 95% 的上限和下限，第一行的 F 统计量显著性值为 0.249 大于 0.05，所以，Triglyceride 和 Final Triglyceride 的相关性不显著。第二行的 F 统计量显著性值为 0.000 小于 0.05，所以，Weight 和 Final Weight 的具有显著的相关性。

| 配对样本检验 | | | | | | | | | |
|--------|-----------------------------------|--------|--------|---------|-------------|--------|--------|-----|----------|
| | | 配对差值 | | | 差值 95% 置信区间 | | t | 自由度 | 显著性 (双尾) |
| | | 平均值 | 标准差 | 标准误差平均值 | 下限 | 上限 | | | |
| 配对 1 | Triglyceride - Final triglyceride | 14.063 | 46.875 | 11.719 | -10.915 | 39.040 | 1.200 | 15 | .249 |
| 配对 2 | Weight - Final weight | 8.063 | 2.886 | .722 | 6.525 | 9.600 | 11.175 | 15 | .000 |

图 6-28 独立样本 t 检验



第 7 章 基本图形的绘制

在常用的统计软件中，绘制的统计图不太美观；而 SPSS 绘制的统计图较为美观，可以满足大多数情况下的要求；STATA 绘制的统计图形最为精美，但由于它采用命令行方式操作，美观的图形需要添加大量选项，普通人不易掌握；而 S-PLUS、MINTAB 等偏数理统计的软件虽然绘图能力也非常强，但由于自身的定位问题，并不为大多数人所熟悉。因此，在各种统计软件中，以 SPSS 制作的统计图应用最为广泛。



本讲内容

- 统计图概述
- 条形图
- 折线图
- 面积图
- 饼图
- 高低图
- 质量控制图
- 箱图
- 散点图
- 直方图
- P-P 图和 Q-Q 图
- 时间序列图

7.1 统计图概述

在统计分析中，统计图作为数据描述的重要方法之一，主要是通过点、线、条、面积等的位置与大小的变化来表现或说明所研究问题的变化及其规律。在数据分析的过程中，数据分析图与数据表格有时可同时产生，有时必须分开进行。

统计图具有简洁、直观、可读性强和易于理解等特点，被分析者和信息使用者广泛使用，因此，数据分析人员进行统计分析时，掌握统计图的绘制与编辑是必不可少的数据

分析技能。在 SPSS 中, 提供了用原始数据和表格中数据进行绘图的功能, 数据图的种类也比较多, 可方便地供数据分析人员选用, 常见统计图可参见表 7-1。

表 7-1 常见统计图

| 符 号 | 图 形 名 称 | 符 号 | 图 形 名 称 | 符 号 | 图 形 名 称 |
|---|---------|---|----------|---|----------|
|  | 条图 |  | 散点图 |  | 折线图 |
|  | 直方图 |  | 饼图 |  | 面积图 |
|  | 箱式图 |  | 正态 Q-Q 图 |  | 正态 P-P 图 |
|  | 质量控制图 |  | Pareto 图 |  | 自回归曲线图 |
|  | 高低图 |  | 交互相关图 |  | 序列图 |
|  | 频谱图 |  | 误差线图 | | |

SPSS 中可以直接利用菜单图形 (Graphs) 来实现, 在本章将对其重点介绍。

7.2 条形图

条形图用条的根数代表分类变量所分组的多少, 或者选用变量的个数, 用条的高度反映各组分析指标值的大小, 或者变量特征值的大小, 各个条之间有间隔。它可以直观揭示或比较频数变量的频数特征值、分类变量在有关综述变量方面的特征值大小, 以此发现重要组或类 (Group)。

下面就利用 SPSS 软件来绘制条形图。

先打开一个 SPSS 数据文件, 选择菜单 “图形 (Graphs) 旧对话框 (Legacy Dialogs) 条形图 (Bars)”, 则弹出如图 7-1 所示对话框, 此对话框用于选择绘制的条形图类型和定义图形中的数据, 对话框组成部分如下。

- 简单条形图 (Simple)。
- 簇状条形图 (Clustered)。
- 堆积条形图 (Stacked)。
- 图表中的数据 (Data in Chart Are): 用于定义图形中的数据的描述方式。个案组摘要 (Summaries for Groups of Cases) 表示观测量分类概述, 对应简单条形图; 单独变量的摘要 (Summaries of Separate Variables) 表示分辨量概述, 对应于分组条形图; 单个个案的值 (Values of Individual Cases) 表示单个观测量值概述。

下面以简单条形图的设置为例。选择图 7-1 中的简单条形图 (Simple), 以及个案组摘要 (Summaries for Groups of Cases) 选项, 然后单击 “定义 (Define)” 按钮, 弹出如图 7-2 所示对话框, 此对话框用于简单条形图的参数设置, 各组成部分具体如下。

- 个案数 (N of Cases): 长条代表记录个数。
- 累计个案数 (Cum.N): 长条代表从前到后的累计记录数。



图 7-1 “条形图 (Bar)” 对话框



图 7-2 “简单条形图的参数设置” 对话框

- 其他统计量 (Other statistic (e.g., mean))：长条代表给定变量的某个统计值，选中此项后则激活其下的变量 (Variable) 框，系统默认长条代表的是变量的均值，如果要自定义则可以单击其下的“更改统计量 (Change Statistic)”按钮来改变，统计量自定义对话框如图 7-3 所示。

1. 条形表示 (Bars Represent) 选项栏

用于定义条形图中长条的具体含义。

- 个案百分比 (% of Cases)：长条代表记录的百分比。
- 累计百分比 (Cum.%)：长条代表从前到后的累计百分比。

2. 类别轴 (Category Axis) 选项

此选项表示分类轴，代表条形图的横坐标，即绘制条形图时的分类变量。

3. 面板划分依据 (Panel by) 选项栏

此栏表示图组变量框，分为行、列两个选项。用于选择变量，按照变量取值不同在同一坐标轴内绘制多张条形图。



图 7-3 “统计量 (Statistic) 自定义” 对话框

4. 模板 (Template) 选项栏

用于选择绘制图形的模板。

5. 标题 (Titles) 设置

单击“标题 (Titles)”按钮,则弹出如图 7-4 所示对话框,此对话框用于设置图形的标题、子标题和脚注。

6. 选项 (Options) 设置

单击“选项 (Options)”按钮,则弹出如图 7-5 所示对话框,此对话框用于定义与缺失值有关的选项。



图 7-4 “标题 (Titles) 设置”对话框



图 7-5 “选项 (Options) 设置”对话框

- 缺失值 (Missing Values) 选项栏: 用于定义对缺失值的处理方式。
- 显示由缺失值定义的组 (Display groups defined by missing values): 选择是否把分类变量的缺失值作为一个组来表示。
- 显示带有个案标签的图表 (Display chart with cases labels): 选择是否把变量值在图中显示为相应点的标签,只有在图中有散点且变量标签存在时此项才可用。
- 误差条形图表示 (Error Bars Represent): 用于设置图形的置信区间、标准差等。

单击图 7-1 中的选项“复式条形图 (Clustered)”图标,然后单击“定义 (Defined)”按钮,则进入作分类条形图的设置界面,与上述的简单条形图设置对话框一致,只是多一个“定义聚类 (Defined Stacked by)”选项框,用于指定一个分类变量。

单击图 7-1 中的选项“堆积条形图 (Stacked)”图标,然后单击“定义 (Defined)”按钮,则进入作分段条形图的设置界面,与上述的简单条形图设置对话框一致,只是多一个“定义聚类 (Defined Stacked by)”选项框,用于指定一个子分类变量。

下面以数据集 CH703.sav 为例来绘制折线图,此数据集是全国各省市、自治区、直辖市的人均生活开支,包括 7 个变量。

下面做变量“衣着”的绘制条形图,选择菜单“图形 (Graphs) 旧对话框 (Legacy

Dialogs) 条形图 (Bars)”, 则弹出如图 7-6 所示对话框, 选中“其他统计量 (Other Statistics)”选项栏, 并选中变量“食品”到“变量 (Variable)”选项栏中, 选中变量“地区”到“类别轴 (Category Axis)”选项栏”中。



图 7-6 “条形图 (Bars) 设置”对话框

设置完成后单击“确定”按钮进行绘制, 各地区的食品均值条形图, 如图 7-7 所示。

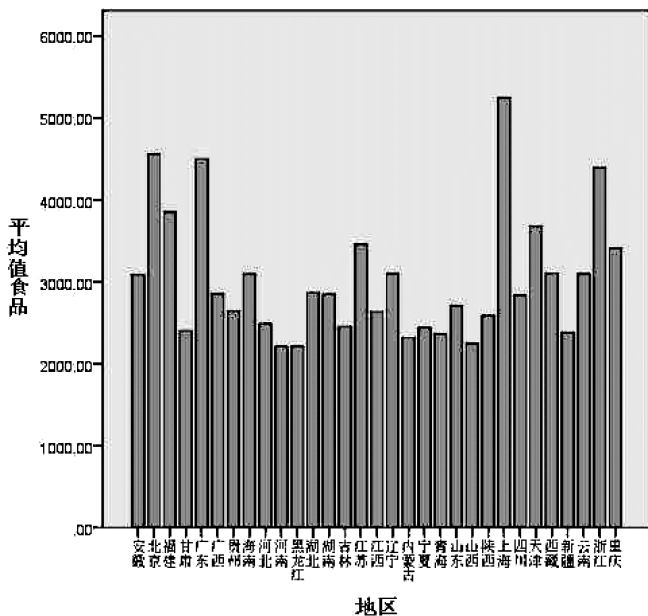


图 7-7 条形图绘制结果

7.3 折线图

折线图用点或折线将各个组别的指标值(或相关变量的成对指标值)连接起来,反映各个组别的指标值的大小(或相关变量的变化趋势),以此发现重要组或重要的变化趋势。折线图和条形图在原理上比较接近,只是外观上有差别,因此两种图形可以转换,在分析过程中,选用哪种图形,与个人的习惯有关,如当分类轴的组别太多时,用条形图就不好看,用折线图表达效果更好。此外,折线图比条形图应用范围更广,如果要反映两个连续变量之间的关系时,只能用折线图来表达。

选择菜单“图形(Graphs) 旧对话框(Legacy Dialogs) 折线图(Line)”,则弹出如图 7-8 所示对话框,此对话框用于选择绘制的条形图类型和定义图形中的数据,对话框组成部分如下。

- 简单线图(Simple)
- 多线线图(Multiple)
- 垂直线图(Drop-line)
- 图表中的数据为(Data in Chart Are):用于定义图形中的数据描述方式。个案组摘要(Summaries for Groups of Cases)表示观测量分类概述,对应简单线图;单独变量的摘要(Summaries of Separate Variables)表示分辨量概述,对应于的线图;单个个案的值(Values of Individual Cases)表示单个观测量值概述。

与简单条形图的设置一致,然后单击图 7-8 中的“定义(Define)”按钮,弹出简单线图的参数设置对话框,此对话框与简单条形图的参数设置一样。

下面以数据集 CH703.sav 为例来绘制折线图,此数据集是全国各省市、自治区、直辖市的人均生活开支,包括 8 个变量。数据集的格式如图 7-9 所示。绘制变量食品消费额的折线图,选择菜单“图形(Graphs) 旧对话框(Legacy Dialogs) 折线图(Line)”,则弹出如图 7-8 所示对话框,单击“定义(Define)”按钮,弹出如图 7-10 所示对话框,选中“其他统计量(Other statistics)”选项栏,并选中变量食品到“变量(Variable)”选项栏中,选中变量地区到“类别轴(Category Axis)”选项栏中。



图 7-8 “折线图”对话框

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|-----------|-----|----|------|----|---|----|----|----|----|----|
| 1 | 地区 | 字符串 | 18 | 0 | | 无 | 无 | 6 | 左 | 名义 | 输入 |
| 2 | 食品 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 3 | 衣着 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 4 | 家庭设备用品及服务 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 5 | 医疗保健 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 6 | 交通和通信 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 7 | 教育文化娱乐服务 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 8 | 居住 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 9 | 杂项商品和服务 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |

图 7-9 数据集格式



图 7-10 “折线图 (Line) 设置”对话框

设置完成以后单击“确定”按钮进行绘制，结果如图 7-11 所示。

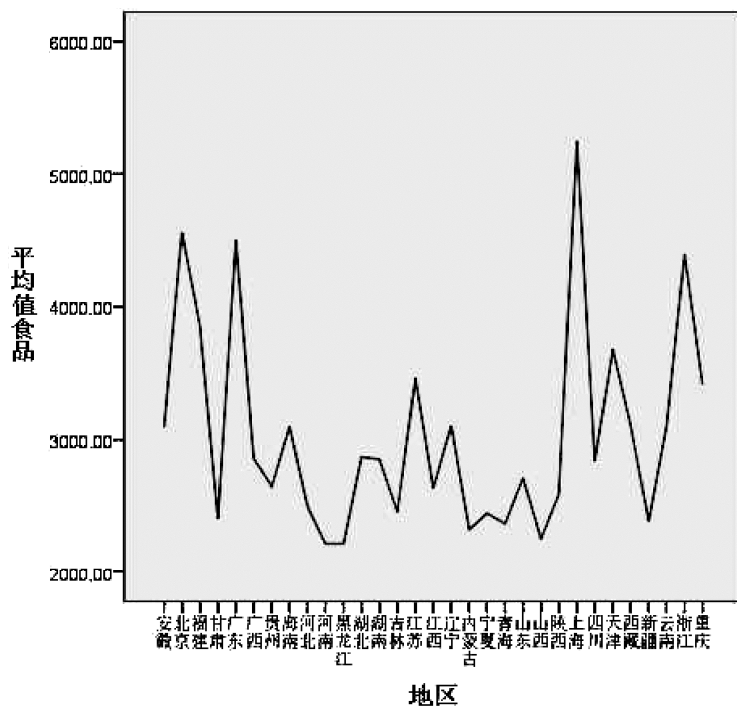


图 7-11 折线图绘制结果

7.4 面积图

条形图、折线图，以及面积图都是用来描述变量的分布情况，且可以相互转换，面积图的参数设置与上述的条形图、折线图的参数设置一致。

选择菜单“图形 (Graphs) 旧对话框 (Legacy Dialogs) 面积图 (Area)”，则弹出如图 7-12 所示对话框，此对话框用于设置面积图的各种参数，对话框组成部分如下。

- 简单箱图 (Simple)
- 堆积面积图 (Stacked)
- 图表中的数据为 (Data in Chart Are): 用于定义图形中的数据描述方式。个案组摘要 (Summaries for Groups of Cases) 表示观测量分类概述，对应简单面积图；单独变量的摘要 (Summaries of Separate Variables) 表示分辨量概述，对应堆积面积图；单个个案的值 (Values of Individual Cases) 表示单个观测量值概述。

与简单条形图的设置一致，然后单击图 7-12 中的“定义 (Define)”按钮，弹出“面积图 (Area) 的参数设置”对话框，此对话框与简单条形图的参数设置一样，如图 7-13 所示。

同样也可以以数据集 CH703.sav 为例来绘制面积图。单击图 7-12 中的“定义 (Define)”按钮，弹出如图 7-13 所示对话框，选中“其他统计量 (Other statistics)”选项栏，并选中变量食品到“变量 (Variable)”选项栏中，选中变量衣着到“类别轴 (Category Axis)”选项栏中。



图 7-12 “面积图 (Area) 参数设置”对话框



图 7-13 “面积图 (Area) 设置”对话框

设置完成以后单击“确定”按钮进行绘制，结果如图 7-14 所示。

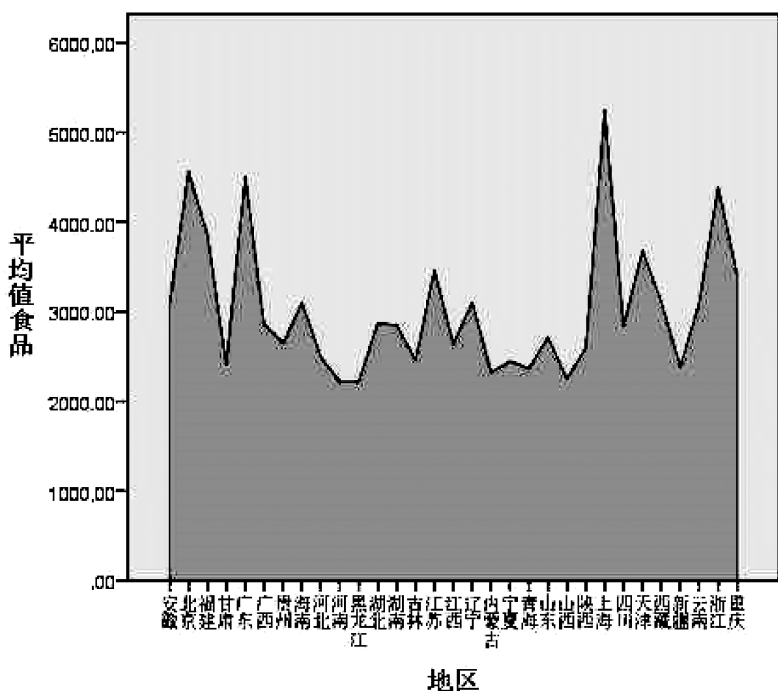


图 7-14 面积图绘制结果

7.5 饼图

与条线图绘制原理相同，饼图类似折线图按照反映指标值的不同分为以下三类：一是描述按照一个频数变量或分类变量分组的各组特征值，简称组特征值饼图或分组饼图。二是描述若干个平行变量的特征值，简称平行变量饼图。三是直接描述原始数据库中的个案数值，简称个案饼图。

7.5.1 饼图参数设置

选择菜单“图形（Graphs）旧对话框（Legacy Dialogs）饼图（Pie）”，则弹出如图 7-15 所示对话框，此对话框用于设置饼图的各种参数，对话框组成部分如下所述。

图表中的数据为（Data in Chart Are）：用于定义图形中的数据的描述方式，各选项含义如下。

- 个案组摘要（Summaries for Groups of Cases）表示观测量分类概述。
- 单独变量摘要（Summaries of Separate Variables）表示分辨量概述。
- 单个个案的值（Values of Individual Cases）表示单个观测量值概述。



图 7-15 “饼图（Pie）参数设置”对话框

与简单条形图的设置一致,然后单击图 7-15 中的“定义(Define)”按钮,弹出饼图的参数设置对话框,此对话框与简单条形图的参数设置一样。



7.5.2 实例分析

以数据集 contacts.sav 为例,数据集格式如图 7-16 所示,绘制出其变量 dept 的饼图分布情况。



| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|------|----|----|------|--------------------|----------------|----|----|----|----|----|
| 1 | dept | 数字 | 4 | 0 | Department | {1, Develop... | 9 | 6 | 右 | 名义 | 输入 |
| 2 | rank | 数字 | 4 | 0 | Company rank | {1, Emplo... | 9 | 6 | 右 | 名义 | 输入 |
| 3 | sale | 数字 | 8 | 2 | Amount of last ... | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 4 | time | 数字 | 4 | 0 | Time since last... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 5 | size | 数字 | 4 | 0 | Size of company | {1, Very sm... | 无 | 6 | 右 | 名义 | 输入 |

图 7-16 数据集格式

-  **结果文件** —— 附带光盘“PROGRAM\CH07\实例 7-1”文件夹
-  **动画演示** —— 附带光盘“AVI\实例 7-1.avi”文件

首先打开数据集,然后选择菜单“图形(Graphs) 旧对话框(Legacy Dialogs) 饼图(Pie)”,则弹出如图 7-15 所示对话框,单击“定义(Define)”按钮,弹出如图 7-17 所示对话框,选择变量 dept 到“定义分区(Define Slices by)”选项栏中,然后单击主界面中的“确定(OK)”按钮进行绘制图形,如图 7-18 所示为所绘制的图形,为变量各个 dept 的所占比例。

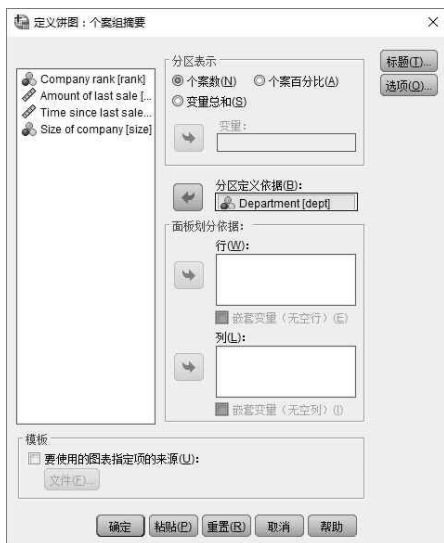


图 7-17 “饼图(Pie)设置”对话框

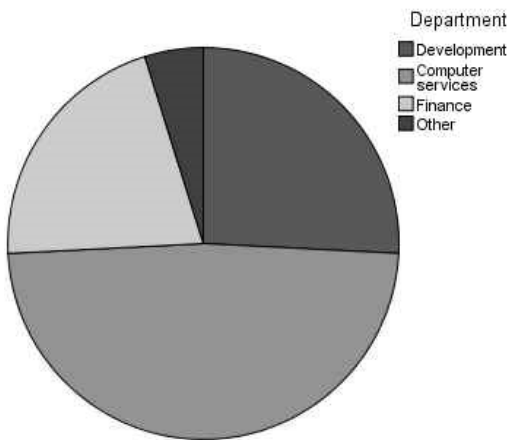


图 7-18 饼图

7.6 高低图

选择菜单“图形 (Graphs) 旧对话框 (Legacy Dialogs) 高低图 (High-Low)”，则弹出如图 7-19 所示对话框，此对话框用于设置高低图的各种参数，对话框组成部分详细如下。

- 简单盘高盘低收盘图 (Simple high-low-close)
- 简单范围条形图 (Simple Range Bar)
- 簇状盘高盘低收盘图 (Clustered high-low-close)
- 簇状范围条形图 (Clustered Range Bar)
- 差别面积图 (Difference Area)
- 图表中的数据为 (Data in Chart Are): 用于定义图 7-19 “高低图参数设置”对话框中的数据的描述方式。个案组摘要 (Summaries for Groups of Cases) 表示观测量分类概述；单独变量的摘要 (Summaries of Separate Variables) 表示分辨量概述；单个个案的值 (Values of Individual Cases) 表示单个观测量值概述。

与简单条形图的设置一致，然后单击图 7-19 中的“定义 (Define)”按钮，弹出高低图的参数设置对话框，此对话框与简单条形图的参数设置一样。

下面就调用高低图过程来绘制高低图，数据集为 stock.sav，数据如图 7-20 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|--------|----|----|------|------|---|----|----|----|----|----|
| 1 | Date | 日期 | 10 | 0 | 日期 | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 2 | Close | 数字 | 8 | 2 | 日收盘的 | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 3 | High | 数字 | 8 | 2 | 日最高的 | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 4 | Low | 数字 | 8 | 2 | 日最低的 | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 5 | Volume | 数字 | 8 | 2 | 成交量 | 无 | 无 | 10 | 右 | 标度 | 输入 |

图 7-20 数据集格式

选择菜单“图形 (Graphs) 旧对话框 (Legacy Dialogs) 高低图 (High-Low)”，则弹出如图 7-19 所示对话框，然后单击“定义 (Define)”按钮弹出如图 7-21 所示对话框，选中变量 high、low、close 到“高 (High)”、“低 (Low)”及“闭合 (Close)”选项栏中。选中变量 date 到“类别标签 (Category Axis)”选项栏中。

设置完成后单击“确定 (OK)”按钮进行绘图，结果如图 7-22 所示。

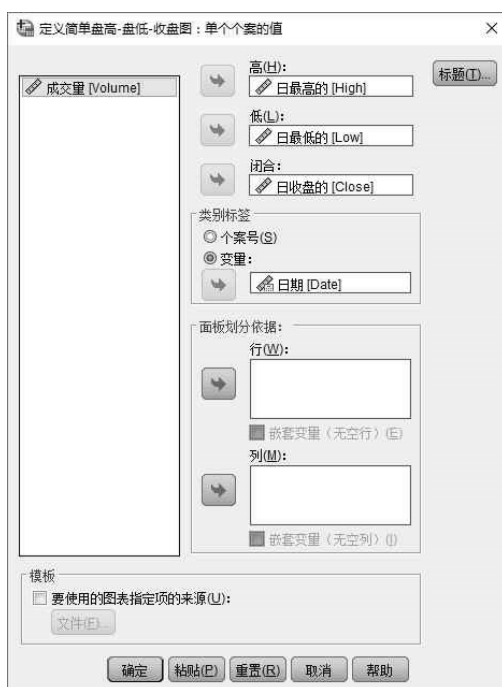


图 7-21 “高低图 (High-Low) 设置”对话框

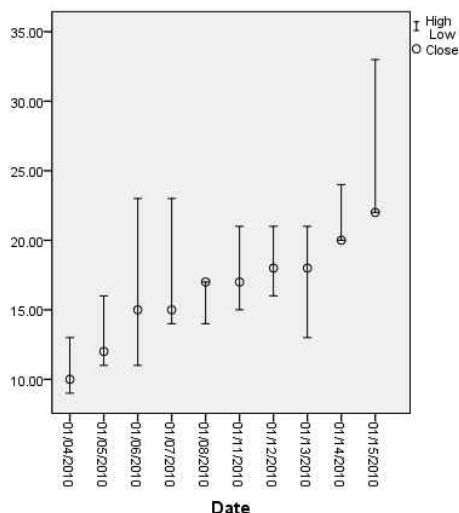


图 7-22 高低图

7.7 质量控制图

将一个时间序列以时间为横轴，在坐标系中依次用点表示出来，并用折线连接起来，同时绘制上限 ULC 和下限 LCL 两条水平控制线及平均值线 Average，以此判断数据的波动是否在控制的范围内。控制图有多种绘制方法，如均值-极差-标准差绘制法 (X-Bar,R,S)、个案-移动极差法 (Individuals,Moving Range)、不合格品率绘制法 (p,np)、缺陷数绘制法 (C,U)。在品质控制中，控制图是最常用的工具。在分析销售数据波动时，也可使用这一工具。

选择菜单“分析 质量控制 (Quality Control) 控制图 (Control Charts)”，则弹出如图 7-23 所示对话框，此对话框用于设置质量控制图的各种参数。对话框组成部分如下所述。

- X 条形图，R 图，S 图：均值与极差组合控制图。
- 个体，移动范围 (Individuals, Moving Range)：单值与移动极差组合控制图。
- p 图，np 图：不合格品率控制图。
- C 图，U 图：缺陷数控制图。
- 数据组织 (Data in Chart Are)：用于定义图形中的数据描述方式。个案是单元 (Cases are units) 表示对观测量作图；个案是子组 (Cases are subgroups) 表示对变量作图。

然后单击如图 7-23 所示的“定义 (Define)”按钮，弹出“质量控制图的参数设置”对话框，此对话框与简单条形图的参数设置一样，如图 7-24 所示。



图 7-23 “质量控制图 (Quality Control) 参数设置”对话框

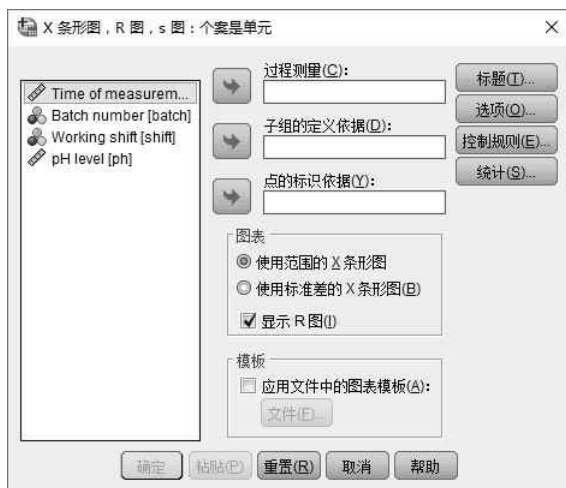


图 7-24 “质量控制图 (Quality Control) 的参数设置”对话框

1. 变量选择

图 7-24 所示的左边为变量列表框。

- 过程测量 (Process Measurement)
- 子组的定义依据 (Subgroups Defined by)
- 点的标识依据 (Identify points by)
- 使用范围的 X 条形图 (X-bar using range): 表示均值-极值图。
- 使用标准差 X 条形图 (X-bar using standard deviation): 表示输出均值-标准差图。
- 显示 R 图 (Display R chart): 表示显示极值或标准差本身的控制图。
- 模板 (Template): 用于指定图形模板, 选中应用图表模板 (Apply chart template from) 后则激活其下的“文件 (File)”按钮, 选择后则制定文件路径。

2. 标题 (Titles) 设置

单击“标题 (Titles)”按钮, 则弹出如图 7-25 所示对话框, 此对话框用于设置图形的标题、子标题和脚注。



图 7-25 “标题 (Titles) 设置”对话框

3. 选项 (Options) 设置

单击“选项 (Options)”按钮,则弹出如图 7-26 所示对话框。

- Sigma 数目 (Number of Sigmas): 指定偏离均值的标准差的范围,系统默认为 3 倍标准差。
- 最小子组大小 (Minimum Subgroup Size): 用于指定每个分组中最少需要的样本数,默认为 2。
- 显示由缺失值定义的子组 (Display subgroups defined by missing values): 表示把缺失值作为一个单独的分组显示于图形中。

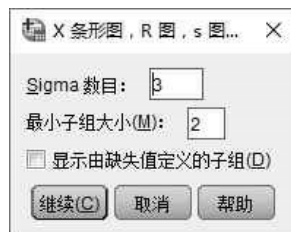


图 7-26 “选项 (Options) 设置”对话框

4. 控制规则 (Control Rules) 设置

单击“控制规则 (Control Rules)”按钮,则弹出如图 7-27 所示对话框,如果某个点违背了此处指定的规则,它将在图中用区别于正常点的形状和颜色表示。

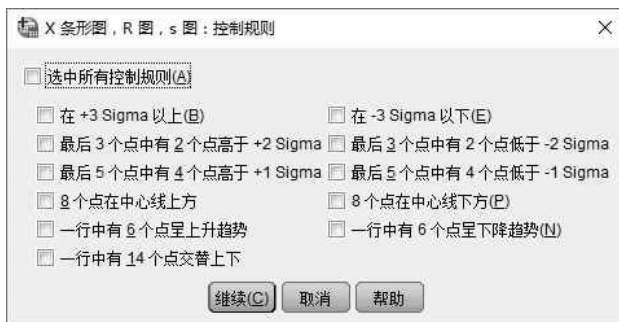


图 7-27 “控制规则 (Control Rules) 设置”对话框

5. 统计量 (Statistics) 设置

单击“统计量 (Statistics)”按钮,则弹出如图 7-28 所示对话框,此对话框用于设置一些统计量。



图 7-28 “统计量 (Statistics) 设置”对话框

- 指定项限制 (Specification Limits): 指定上限 (Upper)、下限 (Lower), 以及一个固定目标值 (Target)。
- 能力 Sigma (Capability Sigma): 指定计算容量 Sigma 时的标准差范围。
- 过程能力指标 (Process Capability Indices): 用于指定衡量工序性能的统计量。
- 过程性能指标 (Process Performance Indices): 用于指定衡量工序性能的统计量, 基本是基于工序的标准差计算得到的。

下面以实例来介绍绘制的操作过程。本实例中使用 SPSS 自带数据集 shampoo_ph.sav, 包含 4 个变量, 即 time、batch、shift、ph, 是关于洗发水的 PH 值的测量数据集, 如图 7-29 所示是数据集的格式。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|-------|----|----|------|-------------------|---------------|----|----|----|----|----|
| 1 | time | 数字 | 4 | 0 | Time of measur... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 2 | batch | 数字 | 4 | 0 | Batch number | 无 | 无 | 7 | 右 | 名义 | 输入 |
| 3 | shift | 数字 | 4 | 0 | Working shift | {1, Night}... | 无 | 7 | 右 | 名义 | 输入 |
| 4 | ph | 数字 | 8 | 2 | pH level | 无 | 无 | 10 | 右 | 标度 | 输入 |

图 7-29 数据集 shampoo_ph.sav 的格式

选择菜单“分析 质量控制 (Quality Control) 控制图 (Control Charts)”, 则弹出如图 7-30 所示对话框, 此对话框用于设置质量控制图的各种参数。选中变量 pH level 到“过程度量 (Process Measurement)”选项栏中, 选择变量 Time of measurement 到“定义子组 (Subgroups Defined by)”选项栏中。



图 7-30 “控制图 (Contral Charts) 设置”对话框

然后单击“统计量 (Statistics)”按钮弹出如图 7-31 所示对话框, 在“上限 (Upper)”、“下限 (Lower)”、“目标 (Target)”选项栏中分别填入 5.4、4.5、5.0。选中 CP、CpU、CpL、K、CpK 和 Z 外选项栏, 选中 PP、PpU、PpL、和 Z 外选项栏, 以及选中“实际%外部规格限制 (Actual % outside specification limits)”选项。最后单击“继续 (continue)”按钮返回主界面。

单击“控制规则 (Control Rules)”按钮弹出如图 7-32 所示的对话框, 选择“选择所有

控制规则 (Select all control rules)”选项栏, 然后单击“继续 (continue)”按钮返回主界面。

设置好上述参数以后单击主界面的“确定”按钮进行绘制图形, 首先是 pH 值的质量控制图形, 如图 7-33、图 7-34 所示, 为均值和范围的变动范围, 同时也会输出一些统计量, 在此不再累述。



图 7-31 “统计量 (Statistics)”对话框

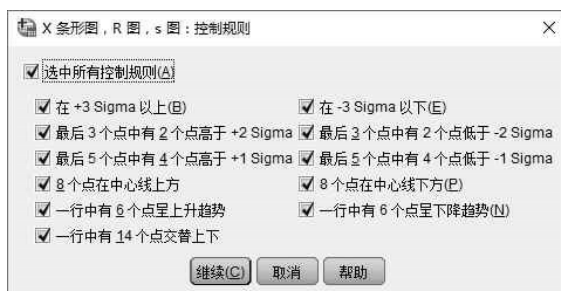


图 7-32 “控制规则设置”对话框

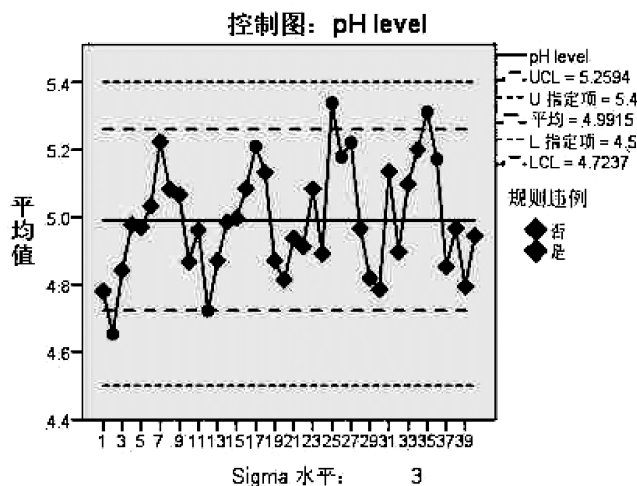


图 7-33 平均值质量控制图

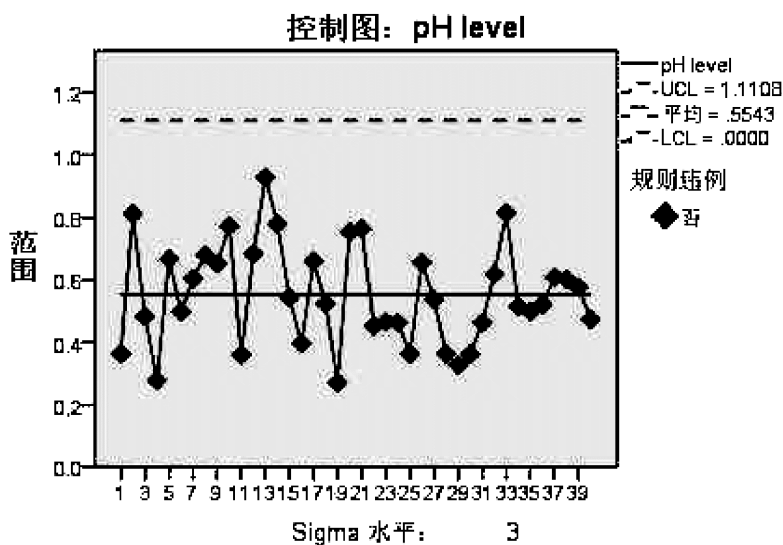


图 7-34 范围质量控制图

7.8 箱图

箱图是一种用于描绘数据分布形式的统计图形，箱图主要包含了指定变量的最小值、1/4 分位图、中位图、3/4 分位图、最大值 5 个统计量。

调用图形 (Graphs) 菜单的箱图 (Boxplot) 过程，可绘制箱图。箱图可用于表现观测数据的中位数、四分位数和两头极端值。

7.8.1 箱图参数设置

选择菜单“图形 (Graphs) 旧对话框 (Legacy Dialogs) 箱图 (Boxplot)”，则弹出如图 7-35 所示对话框，对话框各部分功能如下。

- 简单箱图 (Simple)
- 簇状图 (Clustered)
- 图表中的数据为 (Data in Chart Are): 用于定义图形中的数据描述方式。个案组摘要 (Summaries for groups of cases) 表示观测量分类概述；单独变量的摘要 (Summaries of separate variables) 表示分辨量概述。

单击“定义 (Defined)”按钮，则弹出如图 7-36 所示“箱图设置面板”对话框，各选项栏功能如下。

- 变量列表 (Variable)
- 类别轴 (Category Axis): 用于选入分类变量。
- 标签个案依据 (Label Cases by): 用于选入分类变量。
- 面板依据 (Panel by): 面板变量选择设置。
- 选项 (Options) 按钮的设置与前几节中的方法一致。



图 7-35 “箱图 (Boxplot) 设置”对话框

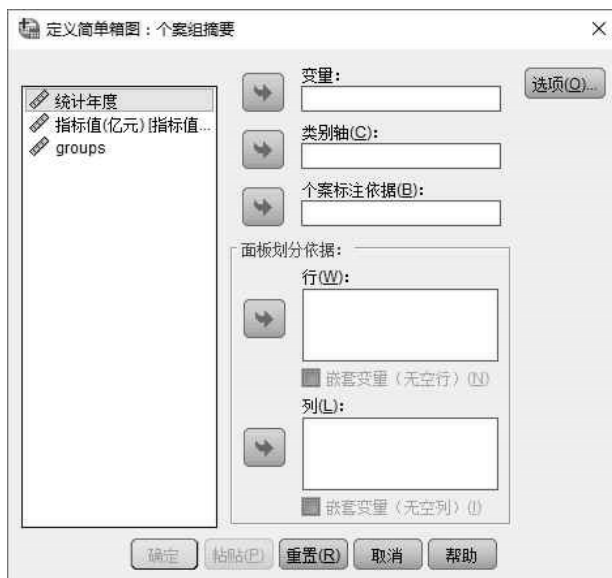


图 7-36 “箱图 (Boxplot) 设置面板”对话框

7.8.2 实例分析

下面以中国国内每年 GDP 的数据为例来绘制时间序列图形，数据集中的 GDP 数据从 1952 年到 2008 年。数据集分为两组，一组是改革开放前的数据；一组是改革开放后的数据，数据集格式如图 7-37 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|--------|----|----|------|---------|-------------|----|----|----|----|----|
| 1 | 统计年度 | 数字 | 11 | 0 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 2 | 指标值亿元 | 数字 | 11 | 0 | 指标值(亿元) | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 3 | groups | 数字 | 8 | 0 | | {1, 改革开放... | 无 | 8 | 右 | 标度 | 输入 |

图 7-37 数据集的格式



结果文件——附带光盘“PROGRAM\CH07\实例 7-2”文件夹



动画演示——附带光盘“AVI\实例 7-2.avi”文件

下面绘制这两组数据的箱图，选择菜单“图形 (Graphs) 旧对话框 (Legacy Dialogs) 箱图 (Boxplot)”，则弹出如图 7-35 所示对话框。然后单击“定义 (Define)”按钮弹出如图 7-38 所示对话框，选中变量指标值到“变量 (Variable)”选项栏中，选中变量 groups 到“类别轴 (Category Axis)”选项栏中，然后单击“确定”按钮进行绘制。

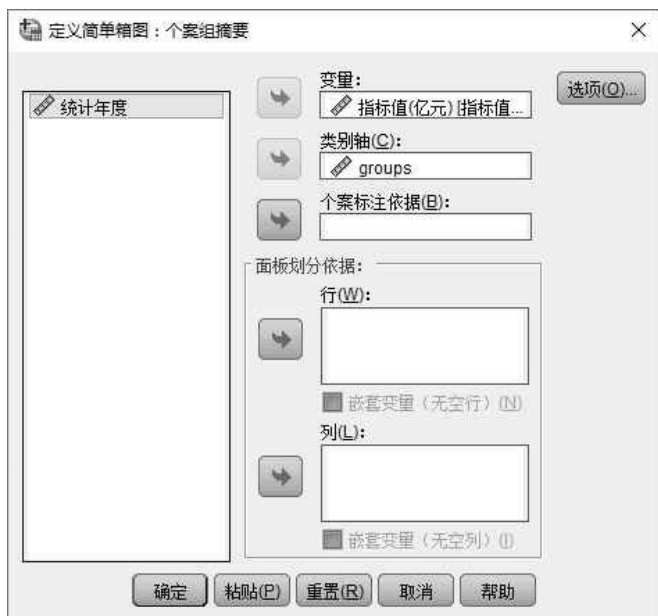


图 7-38 “箱图 (Boxplot) 参数设置”对话框

绘制的箱图如图 7-39 所示，可以看出两组数据的均值等统计信息差别确实是很大，其中在改革开放后的组别中有两个异常点，即 56、57 号的点。

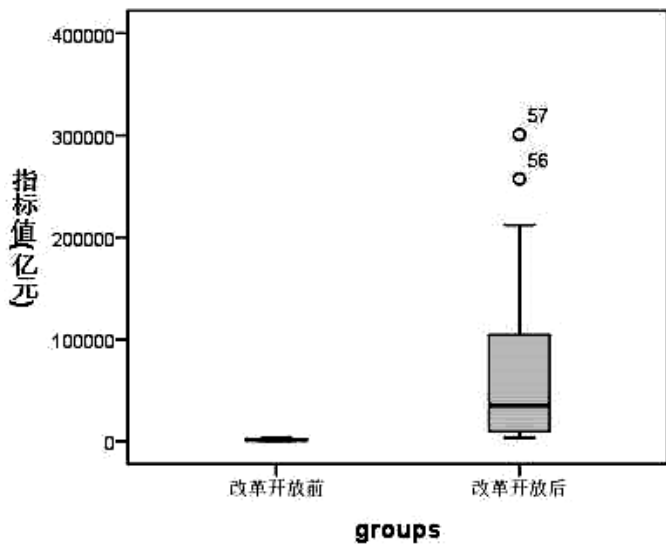


图 7-39 箱图绘制结果

7.9 散点图

散点图用于描绘测量数据的原始分布状态，可以将两个或两个以上变量对应的值在坐标系中用点表示出来，根据点的分布规律或离散程度判断这些变量之间的相关性及其规律。

7.9.1 散点图参数设置

选择菜单“图形 (Graphs) 旧对话框 (Legacy Dialogs) 散点图 / 点图 (Scatterplot)”，则弹出如图 7-40 所示对话框，包含多种散点图。用于设置散点图的各种参数。

选择简单散点图，并单击“定义 (Define)”按钮，则进入如图 7-41 所示的“散点图”对话框。



图 7-40 “散点图/点图 (Scatterplot)”对话框

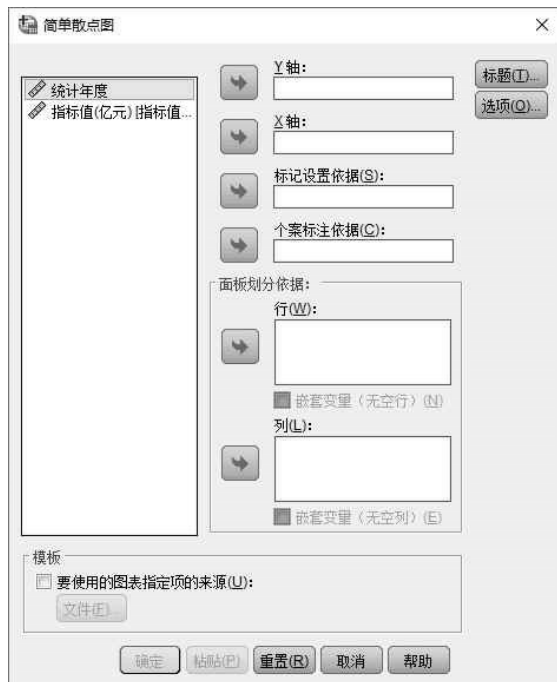


图 7-41 “散点图 (Scatterplot)”对话框

变量选择设置图 7-41 的左边是变量列表，其他各选项框功能如下所述。

- Y 轴。
- X 轴。
- 标记设置依据。
- 个案标注依据。
- 面板划分依据 (Panel Variables)：面板变量设置，包括行和列变量。


模板选项设置：图表规范的使用来源，如果悬着，则激活下面的“文件”按钮。

标题和选项按钮和上述图形的设置一样。

7.9.2 实例分析

下面以中国国内每年 GDP 的数据为例来绘制时间序列图形，数据集集中的 GDP 数据从 1952 年到 2008 年。数据集分为两组，一组是改革开放前的数据；另一组是改革开放后的数据。绘制散点图时将不再分成两组。

 **结果文件** —— 附带光盘 “PROGRAM\CH07\实例 7-3 ” 文件夹

 **动画演示** —— 附带光盘 “AVI\实例 7-3.avi ” 文件

选择菜单 “图形 (Graphs) | 旧对话框 (Legacy Dialogs) | 散点图/点图 (Scatterplot)”, 则弹出如图 7-42 所示对话框, 选中变量指标值到 “Y 轴 (Y Axis)” 选项栏中, 选中变量统计年度到 “X 轴 (X Axis)” 选项栏中, 到此单击 “确定” 按钮即可绘制散点图, 如果用户想标注此图的名称等信息, 也可以单击 “标题 (Title)” 和 “选项 (Options)” 按钮进行设置。



图 7-42 “散点图设置” 对话框

结果如图 7-43 所示, 为数据 GDP 的散点图形。

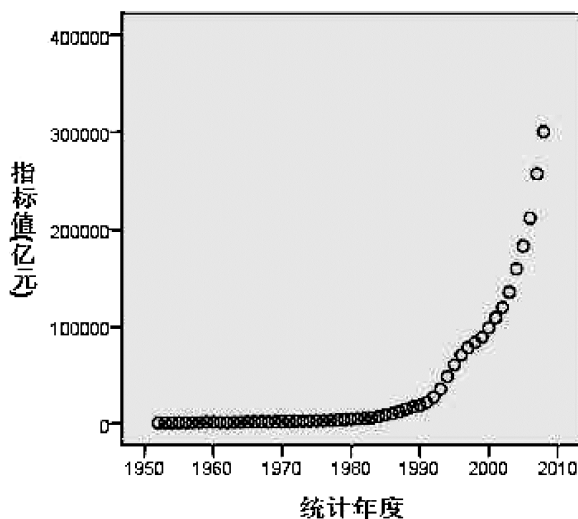


图 7-43 散点图绘制结果

7.10 直方图

直方图用于观察某个变量的分布情况，如果选择了显示正态曲线（Display Normal Curve）复选框，则会同时做出一条当前变量理想状况的正态分布曲线来，和该曲线相比，就可以知道变量的实际分布究竟差了多少。

选择菜单“图形（Graphs）旧对话框（Legacy Dialogs）直方图（Histogram）”，则弹出如图 7-44 所示对话框，用于设置直方图的各种参数。

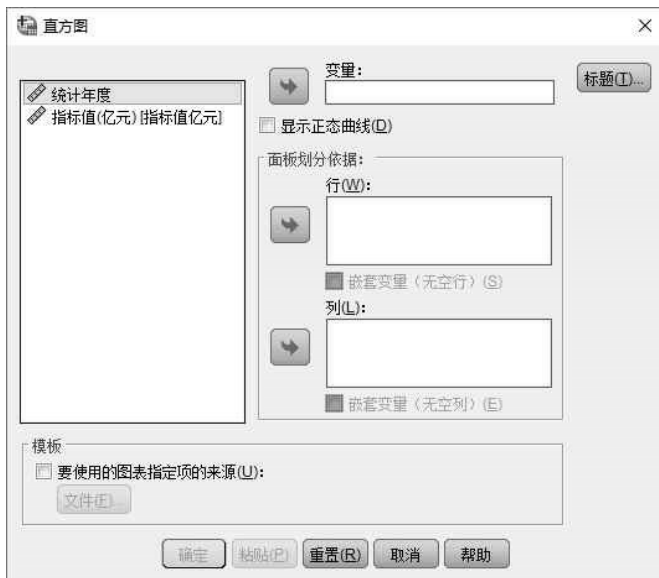


图 7-44 “直方图（Histogram）设置”对话框

图 7-44 的左边是待分析变量列表，其他各组成部分功能如下所述。

变量（Variable）选项栏：此栏用于指定绘制图形的变量，其下的显示正态曲线（Display Normal Curve）复选框表示会同时做出一条当前变量理想状况的正态分布曲线来。

面板划分依据（Panel by）选项栏：此栏用于设置分组变量。

模板（Template）选项栏：此栏用于设置图形模板，其下的“图表规范的使用来源（Use Chart Specifications from）”选项用于指定特殊的图形格式，单击“文件（File）”按钮则打开文件路径。

标题（Titles）选项栏：用于设置图形标题信息，与上述其他类型图形设置方法一致。

在后面的章节中会利用直方图过程来绘制直方图，这里不再介绍。

7.11 P-P 图和 Q-Q 图

P-P 图和 Q-Q 图都是用来观察变量是否服从正态分布的；质量控制图则用来观察个体值是否有超过正常值范围的情况出现。首先介绍 P-P 图。

选择菜单“分析（Analyze）描述统计（Descriptive Statistics）P-P 图（P-P Plots）”，则弹出如图 7-45 所示的对话框。



图 7-45 “P-P 图设置”对话框

变量 (Variables) 选项栏：选择绘制 P-P 图的变量，可以同时选择多个变量。

检验分布 (Test Distribution) 选项栏：用于选择待检测的分布类型，其下的下拉式菜单给出了很多统计分布，例如，Beta 分布、 χ^2 分布、指数分布、伽马分布、半正态分布、Logistic 分布、对数正态分布、正态分布等 13 种分布，其下的 df 用于设置分布自由度。

分布参数选项栏：用于确定分布参数，可由 SPSS 自动从数据中估计，也可以由用户自定义。如选择“从数据中估计 (Estimate from data)”选项，则在 SPSS View 窗口中通过表格输出估计参数值。

转换 (Transform) 选项栏：用于定义数据的转换方式。

- 自然对数转换 (Natural Log Transform)
- 标准值 (Standardize Values)
- 差分 (Difference)
- 季节性差分 (Seasonally Difference)
- 当前周期 (Current Periodicity)

比例估算公式 (Proportion Estimation Formula) 选项栏：用于定义计算预期正态概率值的方法，主要有布洛姆 (Blom's)、秩变换 (Rankit)、图基 (Tukey's)、范德瓦尔登 (Van der Waerden's) 方法。

分配给绑定值的秩 (Rank Assigned to Ties) 选项栏：用于指定对不同的多个变量值的处理方式。包括均值、高、低、强制打开结方法。

Q-Q 图和 P-P 图的定义方式基本一样，二者的区别是 P-P 图比较的是真实数据和待检验分布的累计概率，而 Q-Q 图比较的是真实数据与待检验分布的分位点数。

选择菜单“分析 (Analyze) 描述统计 (Descriptive Statistics) Q-Q 图 (Q-Q Plots)”，则弹出如图 7-46 所示的对话框，此对话框和图 7-45 的 P-P 图的对话框基本相同，在此不再赘述。



图 7-46 “Q-Q 图”对话框

7.12 时间序列图

时间序列图是显示统计数据基本变动规律最简单、最直观的方法，SPSS 软件系统提供了三种时间序列图形的绘制过程。

- 一般时间序列图形（Sequence Charts）
- 交叉相关时间序列图（Cross-correlations）
- 光谱图（Spectral Plot）

7.12.1 时间序列图参数设置

1. 序列图参数设置

选择菜单“分析（Analyze） 时间序列预测 序列图（Sequence Charts）”，则弹出如图 7-47 所示对话框，各部分组成如下所述。

（1）变量选择

图 7-47 的左边是待分析的变量框。

- 变量（Variables）：用于选入作图变量，可以选择多个变量，系统分别绘制图形。
- 时间轴标签（Time Axis Labels）：选入时间轴分类变量。

（2）转换（Transform）选项栏

用于指定作图变量的变换方式。

- 自然对数变换（Natural Log Transform）
- 差分（Difference）
- 季节差分（Seasonally Difference）

- 当前周期 (Current Periodicity)

(3) 每个变量对应一个图表 (One chart per variable)

如果指定了多个绘图变量, 选中此项对每个变量单独输出一张图, 否则所有变量将显示在同一个图里。

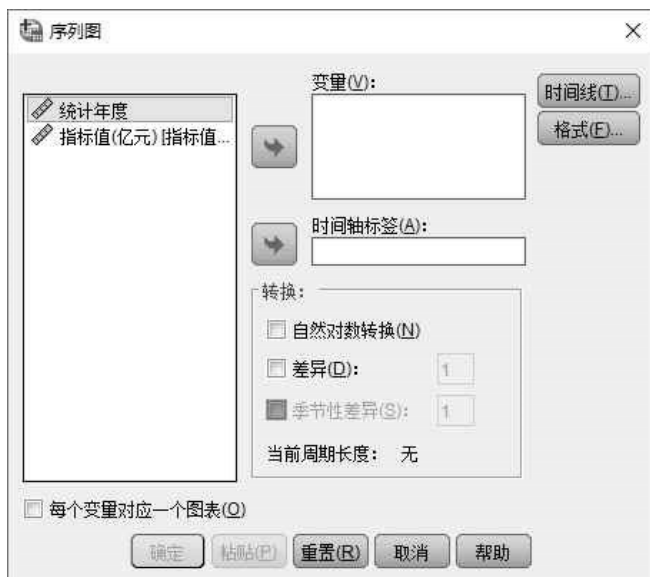


图 7-47 “序列图 (Sequence Charts)”对话框

(4) 时间线 (Time Lines) 设置

单击图 7-47 中的“时间线 (Time Lines)”按钮, 则弹出如图 7-48 所示的对话框, 用于设置关于时间轴参照线的参数。

- 无参考线 (No Reference Lines): 无时间参照线, 系统默认。
- 以下对象的每次变动对应一条参考线 (Lines at Each Change of): 选中后指定参照变量从下面的列表框中选入右侧的参考变量 (Reference Variable) 中, 输出图形里将按照这个变量的取值变化来定义参照线。
- 绘制日期参考线 (Line at date): 表示只显示指定日期的参照线。

(5) 格式 (Format) 设置

选中图 7-47 中的“格式 (Format)”按钮。则弹出如图 7-49 所示对话框, 用于设置图形显示的格式。

- 时间处于水平轴 (Time on Horizontal Axis): 把时间作为横轴, 否则时间轴将显示在纵轴上。
- 单变量图 (Single Variable Chart(s)): 设置关于简单图形的显示选项。包括折线图、面积图、序列平均值的参考线。
- 多变量图 (Multiple Variable Chart): 设置关于复合图形的显示选择项。连接变量之间的个案 (Connect Cases between Variables) 表示把相同时间点上不同序列的取值用线段连接起来。



图 7-48 “时间线 (Time Lines) 设置”对话框



图 7-49 “格式 (Format) 设置”对话框

2. 自相关 (Autocorrelations) 参数设置

选择菜单“分析 (Analyze) 时间序列预测 自相关 (Autocorrelations)”，则弹出如图 7-50 所示对话框。各部分组成如下所述。图中的左边是待分析变量框，“变量 (Variables)”选项框为要进行绘图的变量，从左边的变量框中选出。

输出 (Display) 选项栏：用于选择要输出的图形类型。

- 自相关图形 (Autocorrelations)。
- 偏自相关序列图形 (Partial Autocorrelations)。

转换 (Transform) 选项栏：用于设置变量的转换方式。

- 自然对数变换 (Natural Log Transform)。
- 差异 (Difference)。
- 季节性差异 (Seasonally Difference)。
- 当前周期长度 (Current Periodicity)。



图 7-50 “自相关 (Autocorrelations) 参数设置”对话框

选项 (Options) 设置：单击图 7-50 中的“选项 (Options)”按钮，则弹出如图 7-51 所示对话框，用于设置分析过程中的参数。

- 最大延迟数 (Maximun Number of Lags)：用于指定自相关函数的最大延迟阶数，默认为 16。
- 标准误差法 (Standard Error Method)：用于指定计算标准误差的方法。独立模型 (Independence Model) 表示假设此数据位于白噪声序列；Bartlett 的近似值 (Bartlett's Approximation) 表示计算的标准误差会随着阶数的增加而增加。
- 在周期延迟处显示自相关性 (Display autocorrelations at periodic lags)：表示输出延迟阶数为序列周期长度时的自相关序列。



图 7-51 “选项 Options 设置”对话框

3. 交叉相关性 (Cross-correlations) 参数设置

选择菜单“分析 (Analyze) 时间序列预测 交叉相关性 (Cross-correlations)”，则弹出如图 7-52 所示对话框。各部分组成如下所述。图中的左边是待分析变量框，“变量 (Variables)”选项框为要进行绘图的变量，从左边的变量框中选出。

转换 (Transform) 选项栏：用于设置变量的转换方式。

- 自然对数转换 (Natural Log Transform)
- 差异 (Difference)
- 季节性差异 (Seasonally Difference)
- 当前周期长度 (Current Periodicity)

选项 (Options) 设置：单击图 7-52 中的“选项 (Options)”按钮，则弹出如图 7-53 所示对话框，用于设置分析过程中的参数。



图 7-52 “交叉相关性 (Cross-correlations) 参数设置”对话框

- 最大延迟数 (Maximun Number of Lags): 用于指定自相关函数的最大延迟阶数, 默认为 16。
- 在周期延迟处显示交叉相关性 (Display autocorrelations at periodic lags): 表示输出延迟阶数为序列周期长度时的自相关序列。



图 7-53 “选项 (Options) 设置”对话框

7.12.2 实例分析

下面以中国国内每年 GDP 的数据为例来绘制时间序列图形, 数据集中的 GDP 数据从 1952 年到 2008 年。数据集格式如图 7-54 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|-------|-----|----|------|-------------------|---|----|----|----|----|----|
| 1 | 统计年度 | 数字 | 11 | 0 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 2 | 指标值亿元 | 数字 | 11 | 0 | 指标值(亿元) | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 3 | YEAR_ | 数字 | 8 | 0 | YEAR, not peri... | 无 | 无 | 10 | 右 | 有序 | 输入 |
| 4 | DATE_ | 字符串 | 4 | 0 | Date, Format: ... | 无 | 无 | 7 | 左 | 名义 | 输入 |

图 7-54 数据集的格式



结果文件

——附带光盘“PROGRAM\CH07\实例 7-4”文件夹



动画演示

——附带光盘“AVI\实例 7-4.avi”文件

首先绘制 GDP 的时间序列图形, 选择菜单“分析 (Analyze) 时间序列预测 序列图 (Sequence Charts)”, 则弹出如图 7-55 所示对话框。

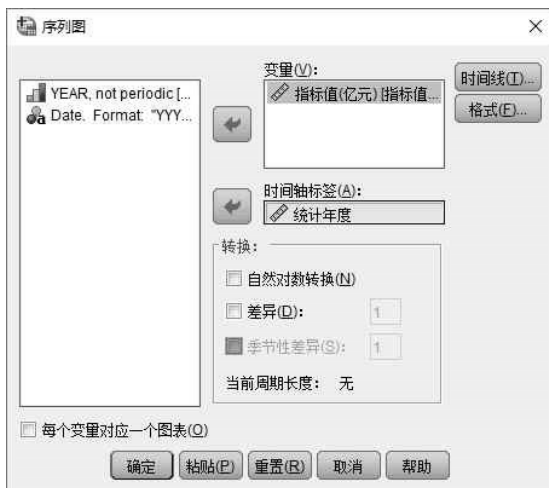


图 7-55 “序列图 (Sequence Charts)”对话框

选择如图 7-55 所示的变量，其他选择项默认，然后单击“确定”按钮进行绘制，结果如图 7-56 所示。图中可以看出，改革开放以后中国的 GDP 数据上升很快。

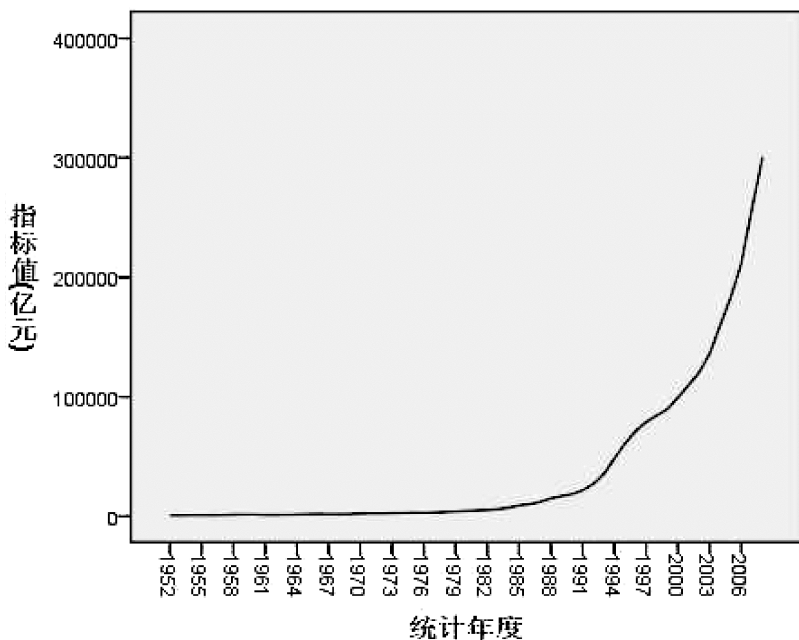


图 7-56 绘制结果

当然，也可以绘制自相关序列图，选择菜单“分析（Analyze） 时间序列预测 自相关（Autocorrelations）”，则弹出如图 7-57 所示对话框。选中变量指标值到“变量（Variables）”选项栏中。然后单击“确定”按钮进行绘制。



图 7-57 参数设置

结果如图 7-58 所示，是自相关系数图，显然有较明显的拖尾现象。

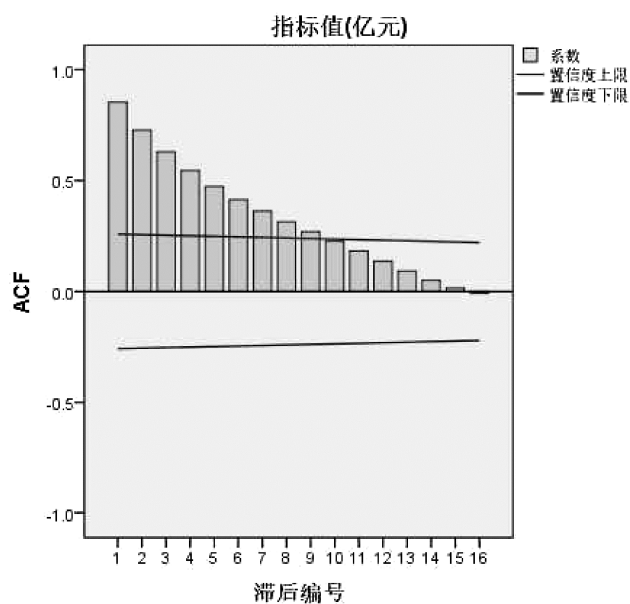


图 7-58 自相关系数图

最后是偏相关系数图，如图 7-59 所示，可以用于进行时间序列模型的分析。

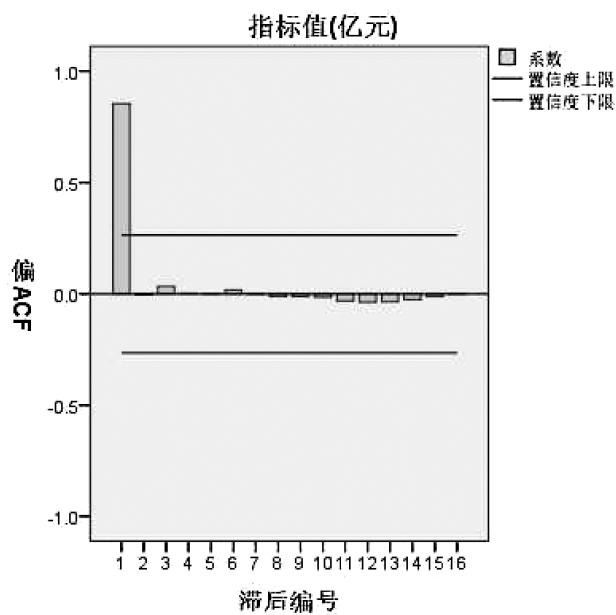
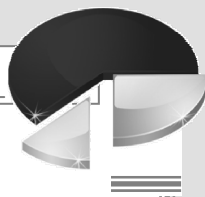


图 7-59 偏相关系数图



第 8 章 非参数检验

本章讲述非参数检验方法，与第 6 章所讲的检验不同，使用这类方法不需要对总体分布做任何事先的假设（如正态总体）。同时从检验的内容来说，也不是检验总体分布的某些参数（如均值、成数、方差等），而是检验总体某些有关的性质，所以称为非参数检验。非参数检验，泛指“对分布类型已知的总体进行参数检验”之外的所有检验方法。

非参数检验内容很多，本章介绍常用的 χ^2 检验、符号检验（Sign Test）、秩和检验（Rank-Sum Test）、K-S 检验等非参数检验方法。



本讲内容

- 非参数检验概述
- χ^2 检验
- 二项分布检验
- 游程检验
- 单样本 K-S 检验
- 两独立样本分布位置检验
- 多个独立样本分布位置检验
- 两相关样本分布位置检验
- 多个独立样本分布位置检验

8.1 非参数检验概述

非参数检验法从实质上讲，只是检验总体分布的位置（中位数）是否相同，所以，对于总体分布已知的样本也可以采用非参数检验法，但是由于它不能充分利用样本内所有的数量信息，检验的效率一般要低于参数检验方法。

与均值差等检验比较，非参数检验有什么优点呢？在对均值差进行 t 检验时，不仅要有定距尺度的假设，还要有正态总体的假设。当然，对于大样本，正态总体的假设可以放松。但正是对于小样本，这种假设最容易出问题。因此，在满足下面两条件之一时，期望用非参数检验代替均值差检验。

- 没有采用定距尺度，但可以安排数据的顺序（秩）。
- 样本小且不能假设具有正态分布。

由于非参数检验不能充分利用全部现有的资料信息。因此，如果有根据采用定距尺度，并且如果对于小样本能够假设其具有正态性，或对于大样本能够放松对正态性假设的要求，一般宁愿使用均值差检验，而不用非参数检验。

非参数检验，无须做出经典统计所必要的关于分布的任何假设。唯一需要的假设是，全部数据或数据对都出自相同的基本总体，且取样是随机的、相互独立的。基于这种原因，非参数检验又称为分布自由（或无分布）检验。“无分布”不是指总体真的无分布，而是指虽有时对总体分布一无所知，但仍可以进行分析。不仅如此，这些很容易理解的方法还可以用于处理等级的资料和定性的信息。

非参数检验法与参数检验法相比，特点可以归纳如下。

- 非参数检验一般不需要严格的前提假设。
- 非参数检验特别适用于顺序资料。
- 非参数检验很适用于小样本，并且计算简单。
- 非参数检验法最大的不足是没能充分利用数据资料的全部信息。
- 非参数检验法目前还不能用于处理因素间的交互作用。

非参数检验的方法很多，分别适用于各种特点的资料。本节将介绍几种常用的非参数检验方法。

参数检验和非参数检验之间的详细比较，可以参见表 8-1。

表 8-1 参数检验和非参数检验

| 检 验 类 型 | 参 数 | 非 参 数 |
|---------|----------|---------------------|
| 单样本 | z 和 t 检验 | 符号检验 |
| | | Wilcoxon 符号秩检验 |
| 两独立样本 | z 和 t 检验 | Wilcoxon 秩和检验 |
| | | Mann Whitney 的 U 检验 |
| 几个独立的样本 | CRD 方差分析 | Kruskal-Wallace 检验 |
| 几个匹配的样本 | RBD 方差分析 | Friedman 检验 |
| 相关性 | 皮尔森 | Spearman 秩相关 |
| | | Kendall 的秩相关 |

8.2 χ^2 检验

χ^2 拟合优度检验（Chi-square Goodness-of-fit Test）适用于具有明显分类特征的数据。如要研究消费者对某种产品是否有“颜色”的偏好，可以将 200 位消费者按购买不同颜色的产品分类，得到各颜色购买者的人数。根据这些样本数据来判断样本所属的总体分布与某一设定分布是否有显著差异，设定分布可以是熟悉的理论分布，如正态分布、均匀

分布等，也可以是任何想象的分布。零假设 H_0 是样本所属总体其分布形态与设定分布无显著差异。在进行检验时需要构造下面的 χ^2 统计量，即

$$\chi^2 = \sum_{i=1}^k \frac{(f_{0i} - f_{ei})^2}{f_{ei}}$$

式中： k 是样本分类的个数； f_{0i} 表示实际观察到的频率； f_{ei} 表示设定频率，即理论频率。

可见，如果观察频率与设定频率越接近，则 χ^2 值越小，根据皮尔逊定理，当 n 充分大时， χ^2 统计量渐近服从于 $k-1$ 个自由度的 χ^2 分布。可以计算出 χ^2 统计量，判断有两种方法。

一是依据 χ^2 分布表，给出所对应的概率值，如果该概率值小于或等于给定的显著性水平 α ，则拒绝 H_0 ，即样本所属的总体分布形态与设定的分布存在显著差异；如果该概率值大于给定的显著性水平 α ，则不能拒绝 H_0 ，即没有理由认为样本所属的总体分布形态与设定分布有显著差异。

二是依据 χ^2 分布表，给出 α 所对应的临界值 χ_α^2 ，如果 χ^2 统计量大于或等于临界值，则拒绝 H_0 ，认为样本所属的总体分布形态与设定分布存在显著差异；如果 χ^2 统计量小于临界值，则不能拒绝 H_0 。

由于奠定检验基础的皮尔逊定理要求样本是充分大，所以，在搜集资料时必须要保证样本容量不小于 50，同时每个单元中的期望频率不能太小，如果第一次分类时有单元中的频率 $f_e < 5$ ，则需要将它与相邻的组进行合并，如果 20% 的单元理论频率 $f_e < 5$ ，则不能用 χ^2 检验了。

8.2.1 χ^2 检验的参数设置

选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框 卡方检验 (Chi-square Test)”，则弹出如图 8-1 所示对话框，此对话框用于设置 χ^2 检验的各种参数。



图 8-1 “ χ^2 检验设置”对话框

1. 变量设置

图 8-1 中的左上角是待分析变量框。

检验变量列表 (Test Variable List) 选项栏：用于从列表框中选出检验变量，必须是数值型分类变量，若选入多个变量，则分别处理。

期望范围 (Expected Range) 选项栏：用于设置检验变量取值的区间范围。

- 从数据中获取 (Get from Data)：表示检验变量每个唯一的取值都作为一个类别，为系统默认选项。
- 使用指定范围 (Use Specified Range)：用户自定义，选中后则激活其下的“下限 (Lower)”、“上限 (Upper)”选择框。

期望值 (Expected Values) 选项栏：用于设置待检验理论期望值的具体取值。

- 所有类别相等 (All Categories Equal)：表示每个类别的期望取值都相等。
- 值 (Values)：用户自定义的期望值，输入期望值以后则激活其下的“添加”、“更改”、“删除”按钮。

2. 精确 (Exact) 设置

如图 8-1 所示，单击“精确 (Exact)”按钮，则弹出如图 8-2 所示对话框。各功能介绍如下。

- 仅渐进法 (Asymptotic only)：只计算近似概率。
- 蒙特卡洛法 (Monte Carlo)：利用蒙特卡洛法方法计算精确概率，其下选项框用于自行设置置信区间和样本数。
- 精确 (Exact)：在给定时间内计算精确概率值，如果超过给定时间则停止计算。

3. 选项 (Options) 设置

如图 8-1 所示，单击“选项 (Options)”按钮，则弹出如图 8-3 所示对话框。各功能含义如下。

- 统计量 (Statistics)：输出统计量。描述 (Descriptive) 表示描述性统计量；还有四分位数 (Quartiles)。



图 8-2 “精确 (Exact) 设置”对话框



图 8-3 “选项 (Options) 设置”对话框

- 缺失值 (Missing Values): 缺失值处理方法。按检验排除个案 (Export Cases Test-by-test) 表示对每一个检验变量来个别地排除缺失值; 成列表排除个案 (Export Cases Listwise) 表示凡含有缺失值的观测量全部从分析中排除。

8.2.2 χ^2 检验实例分析

本实例所用 SPSS 自带的数据文件 dischargedata.sav, 此数据集包含四个变量 dow、day、discharg、admits, 下面对此数据集进行 χ^2 检验。此数据集的格式如图 8-4 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|----------|-----|----|------|-------------------|----------------|----|---|----|----|----|
| 1 | dow | 数字 | 2 | 0 | Day of the Week | {1, Sunday}... | 无 | 8 | 右 | 标度 | 输入 |
| 2 | day | 字符串 | 3 | 0 | Week Day Name | 无 | 无 | 8 | 左 | 名义 | 输入 |
| 3 | discharg | 数字 | 3 | 0 | Average Daily ... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 4 | admits | 数字 | 3 | 0 | Average Daily ... | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 8-4 数据集 dischargedata.sav 的数据格式

- **结果文件** —— 附带光盘 “PROGRAM\CH08\实例 8-1” 文件夹
- **动画演示** —— 附带光盘 “AVI\实例 8-1.avi” 文件

1. 变量设置

首先对原始数据集进行预处理, 选择菜单 “数据 (Data) 个案加权 (Weight Cases)”, 弹出如图 8-5 所示对话框。选择 “加权个案 (Weight cases by)” 选项, 然后选择变量 Average Daily Discharges 到 “频率变量 (Frequency Variable)” 选项栏中, 然后单击 “确定” 按钮。



图 8-5 “加权 (Weight Cases) 变量设置” 对话框

然后选择菜单 “分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框 卡方检验 (Chi-square test)”, 则弹出如图 8-6 所示对话框, 此对话框用于设置 χ^2 检验的各种参数。选择 Day of the Week 变量到 “检验变量列表 (Test Variable List)” 选项栏中。

图 8-6 “ χ^2 检验 (Chi-square test) 设置”对话框

2. 结果分析

单击“确定 (OK)”按钮, 则进行 χ^2 检验分析, 如图 8-7 所示为基本统计信息, 包含实测个案数、期望个案数及残差。

| Day of the Week | | | |
|-----------------|-------|-------|-------|
| | 实测个案数 | 期望个案数 | 残差 |
| Sunday | 44 | 84.1 | -40.1 |
| Monday | 78 | 84.1 | -6.1 |
| Tuesday | 90 | 84.1 | 5.9 |
| Wednesday | 94 | 84.1 | 9.9 |
| Thursday | 89 | 84.1 | 4.9 |
| Friday | 110 | 84.1 | 25.9 |
| Saturday | 84 | 84.1 | -.1 |
| 总计 | 589 | | |

图 8-7 基本统计信息

如图 8-8 所示的是检验统计量。 χ^2 统计量为 29.389, 渐进显著性的取值 0.000 小于 0.01, 所以, 在 0.01 的显著性水平上否定零假设。

| 检验统计 | |
|--|---------------------|
| Day of the Week | |
| 卡方 | 29.389 ^a |
| 自由度 | 6 |
| 渐进显著性 | .000 |
| a. 0 个单元格 (0.0%) 的期望频率低于 5。期望的最低单元格频率为 84.1。 | |

图 8-8 检验统计量

8.3 二项分布检验

8.3.1 二项分布检验的参数设置

选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框 二项式 (Binomial Test)”，则弹出如图 8-9 所示对话框，对话框主要组成部分如下所述。

变量选择：如图 8-9 中左上角为待变量列表。

- 检验变量列表 (Test Variable List)：检验变量选项框，用于放置检验变量，可以选择多个检验变量。
- 定义二分法 (Define Dichotomy)：定义二元变量选项框，当检验变量已经是二元变量时，选择“从数据中获取 (Get from Data)”选项；当检验变量不是二元变量时，则选择“分割点 (Cut point)”选项，并在其后的空白栏内输入断点值。
- 检验比例 (Test Proportion)：用于指定检验的零假设。系统默认为 0.50。

精确 (Exact) 设置：与 χ^2 检验设置对话框一致。

选项 (Options) 设置：与 χ^2 检验设置对话框一致。



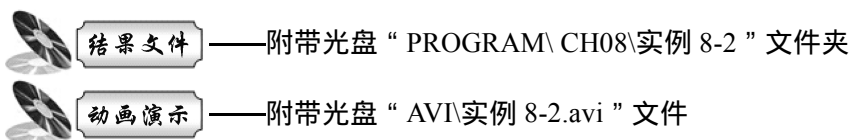
图 8-9 “二项分布 (Binomial Test) 检验参数设置”对话框

8.3.2 实例分析

本实例所用 SPSS 自带的数据文件 telco.sav，此数据集包含 42 个变量，下面对此数据集进行二项分布检验，此数据集的格式如图 8-10 所示。

| telco.sav [数据集1] - IBM SPSS Statistics 数据编辑器 | | | | | | | | | | | |
|--|---------|----|----|------|--------------------|-----------------|----|----|----|----|----|
| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
| 1 | region | 数字 | 4 | 0 | Geographic indi... | {1, Zone 1}... | 无 | 6 | 右 | 名义 | 输入 |
| 2 | tenure | 数字 | 4 | 0 | Months with se... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 3 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 4 | marital | 数字 | 4 | 0 | Marital status | {0, Unmarrie... | 无 | 7 | 右 | 名义 | 输入 |
| 5 | address | 数字 | 4 | 0 | Years at curren... | 无 | 无 | 7 | 右 | 标度 | 输入 |
| 6 | income | 数字 | 8 | 2 | Household inco... | 无 | 无 | 10 | 右 | 标度 | 输入 |

图 8-10 数据集 dischargedata.sav 的数据格式



结果文件——附带光盘“PROGRAM\CH08\实例 8-2”文件夹

动画演示——附带光盘“AVI\实例 8-2.avi”文件

1. 变量设置

首先对原始数据集进行预处理, 选择菜单“数据(Data) 拆分文件(Split File)”, 弹出如图 8-11 所示对话框。选择“按组组织输出(Compare Groups)”选项, 然后选择变量 Customer category 到“分组依据(Groups Based on)”选项栏中, 单击“确定”按钮。



图 8-11 “分割文件(Split File)设置”对话框

然后选择菜单“分析(Analyze) 非参数检验(Nonparametric Tests) 旧对话框 二项式(Binomial)”, 则弹出如图 8-12 所示对话框, 此对话框用于设置二项分布检验的各种参数。选择 Churn within last month 变量到“检验变量列表(Test Variable List)”选项栏中, 在其下的“检验比例(Test Proportion)”选项栏中填写 0.27。

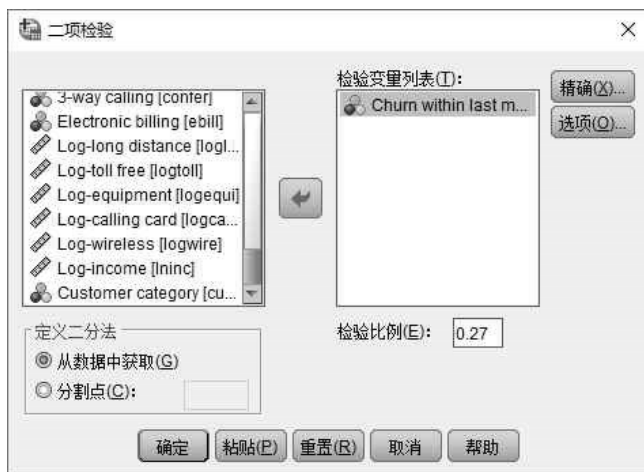


图 8-12 “二项式检验(Binomial)设置”对话框

然后单击“选项 (Options)”按钮，则弹出如图 8-13 所示对话框，选择“统计量中的描述性 (Descriptive)”选项栏，然后单击“继续 (Continue)”按钮返回主界面。

2. 结果分析

单击二项式检验主界面的“确定”按钮，则进行二项分布检验分析，如图 8-14 所示为基本描述性统计信息。包含观察数、均值，以及标准差等信息。



图 8-13 “选项 (Options) 设置”对话框

| 描述统计 ^a | | | | | |
|--------------------------------------|-----|-----|------|-----|-----|
| | 个案数 | 平均值 | 标准差 | 最小值 | 最大值 |
| Churn within last month | 236 | .37 | .485 | 0 | 1 |
| a. Customer category = Total service | | | | | |

图 8-14 基本统计信息

如图 8-15 所示的是二项分布检验统计量。统计量为 0.000 远远小于 0.10，所以否定零假设。

| 二项检验 ^a | | | | | | |
|--------------------------------------|-----|-----|-----|------|------|---------------|
| | | 类别 | 个案数 | 实测比例 | 检验比例 | 精确显著性 (单尾) |
| Churn within last month | 组 1 | Yes | 88 | .37 | .27 | .000 |
| | 组 2 | No | 148 | .63 | | |
| | 总计 | | 236 | 1.00 | | |
| a. Customer category = Total service | | | | | | |

图 8-15 检验统计量

8.4 游程检验

游程检验是适用于独立样本的另一种检验法。游程检验的基本原理和计算方法很简单：先把两个样本混合起来，按大小排列，并赋予其秩。那么，当样本所属的总体是同分布的话，不太可能出现来自总体 1 的样本全是高秩，而来自总体 2 的样本全是低秩的情况；反之亦然。可能性最多的情况是，来自总体 1 和总体 2 的样本，其秩是随机交错的。因此，根据混合样本中两样本交错的次数来检定秩交错次数是随机的零假设，这就是游程检验，其具体步骤如下。

- 设从两个未知的总体 1 和总体 2 中分别独立、随机地各抽取一个样本，样本 1 的容量为 n_1 ，样本 2 的容量为 n_2 。
- 把样本 1 和样本 2 混合起来，并按数值从小到大顺序编号，每个数据的编号就是它的秩。
- 点算游程数目。一个游程指混合样本中接连属于一个样本的一串秩，其前后是另一个样本的秩。根据显著性水平 α 确定否定域 $r_\alpha(n_1, n_2)$ ，游程数目 r 的抽样分布可用

于建立否定零假设的否定域。

- 检定零假设。以混合样本中的游程数目 r 为检验统计量，如果游程的数目很大，就表明两个样本混合得很好，不能否定零假设；相反，如果游程的数目较小，零假设就很可能是错的，应该否定。

8.4.1 游程检验的参数设置

选择菜单“分析（Analyze） 非参数检验（Nonparametric Tests） 旧对话框 游程（Runs test）”，则弹出如图 8-16 所示对话框，对话框主要组成部分如下所述。

变量选择：如图 8-16 中左上角为待变量列表。

- 检验变量列表（Test Variable List）：检验变量选项框，用于放置检验变量，可以选入多个检验变量。
- 分割点（Cut Point）：断点设置选项，系统提供了 4 种方法，分别是中位数（Median）、平均值（Mean）、众数（Mode）、定制（Custom）。

精确（Exact）设置：与 χ^2 检验设置对话框一致。

选项（Options）设置：与 χ^2 检验设置对话框一致。



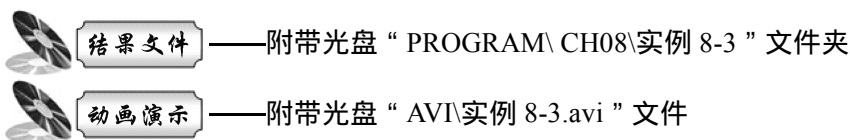
图 8-16 “游程（Runs test）检验参数设置”对话框

8.4.2 实例分析

两种操作方法对劳动效率的影响，随机抽取 12 人用第一种操作方法。10 人用第二种操作方法，每人的日产量参见表 8-2，两种操作方法有无显著差异。

表 8-2 产量表

| 序号 | 第一组产量 | 第二组产量 | 序号 | 第一组产量 | 第二组产量 |
|----|-------|-------|----|-------|-------|
| 1 | 55 | 65 | 7 | 73 | 86 |
| 2 | 59 | 77 | 8 | 75 | 91 |
| 3 | 61 | 80 | 9 | 76 | 91 |
| 4 | 64 | 80 | 10 | 82 | 92 |
| 5 | 64 | 84 | 11 | 82 | |
| 6 | 70 | 84 | 12 | 83 | |



建立假设如下。

H_0 : 两种操作方法没有显著差异;

H_1 : 两种操作方法的差异是显著的。

1. 参数设置

选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框→游程 (Runs test)”, 打开“游程检验 (Runs Test)”对话框, 如图 8-17 所示。

将变量组别进入“检验变量列表”选项框中, 在“分割点 (Cut Point)”选择栏中选择划分二类的检验分类点, 系统默认中位数。本例中选择平均值作为检验分类点, 在选项 (Options) 框内的统计量中勾选描述性。



图 8-17 “游程检验 (Runs Test)”对话框

2. 结果分析

设置好以后单击图 8-17 中的“确定 (OK)”按钮进行分析, 首先是如图 8-18 所示的描述性统计量, 包括观测数、均值方差等信息。

| 描述统计 | | | | | |
|------|-----|------|------|-----|-----|
| | 个案数 | 平均值 | 标准差 | 最小值 | 最大值 |
| 组别 | 22 | 1.45 | .510 | 1 | 2 |

图 8-18 描述性统计结果

然后是游程检验结果, 如图 8-19 所示。按照产量排序后, 组别标志值的游程为 2, 由样本计算的检验统计量 Z 为 -4.147, P 值为 0.000, 小于 0.05, 拒绝原假设 H_0 , 即认为两种操作方法的差异显著。

| 游程检验 2 | |
|------------------|--------|
| 组别 | |
| 检验值 ^a | 1.45 |
| 个案数 < 检验值 | 12 |
| 个案数 ≥ 检验值 | 10 |
| 总个案数 | 22 |
| 游程数 | 2 |
| Z | -4.147 |
| 渐近显著性(双尾) | .000 |
| a. 平均值 | |

图 8-19 游程检验结果

8.5 单样本 K-S 检验

单样本 K-S 检验 (1-Sample K-S test) 是以两位苏联数学家柯尔莫哥 (Kolmogorov) 和斯米诺夫 (Smirnov) 命名的。K-S 检验是一种拟合优度检验, 研究样本观察值的分布和设定的理论分布间是否吻合, 通过对两个分布差异的分析确定是否有理由认为样本的观察结果来自所设定的理论分布总体。

设 $S_n(x)$ 是一个 n 次观察的随机样本观察值的累积概率分布函数, 即经验分布函数; $F_0(x)$ 是一个特定的累积概率分布函数, 即理论分布函数。定义 $D = |S_n(x) - F_0(x)|$, 显然若对每一个 x 值来说, $S_n(x)$ 与 $F_0(x)$ 十分接近, 也就是差异很小, 则表明经验分布函数与理论分布函数的拟合程度很高, 有理由认为样本数据来自具有该理论分布的总体。K-S 检验主要考察的是绝对差数 $D = |S_n(x) - F_0(x)|$ 中那个最大的偏差, 即利用下面的统计量作出判断。

$$D_{\max} = \max |S_n(x) - F_0(x)|$$

K-S 检验的步骤如下。

提出假设: $H_0: S_n(x) = F_0(x)$, $H_1: S_n(x) \neq F_0(x)$ 。

计算各个 D , 找出统计量 D_{\max} 。

查找临界值: 根据给定的显著性水平 α 和样本数据个数 n , 查《单样本 K-S 检验统计量表》可以得到临界值 D_α 。

做出判定: 若 $D_{\max} > D_\alpha$, 则在 α 水平上, 拒绝 H_0 ; 若 $D_{\max} < D_\alpha$, 则不能拒绝 H_0 。

8.5.1 单样本 K-S 检验的参数设置

选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框 单样本 K-S 检验 (1 Sample K-S)”, 则弹出如图 8-20 所示对话框, 对话框主要组成部分如下所述。

变量选择: 如图 8-20 中左上角为待变量列表。

- 检验变量列表 (Test Variable List): 检验变量选项框, 用于放置检验变量, 可以选入多个检验变量。
- 检验分布 (Test Distribution): 检验的概率分布选项框, 系统提供了四种常见的方法, 分别是常规分布 (Normal) 泊松分布 (Poisson) 相等分布 (Uniform) 指数

分布 (Exponential)。

精确 (Exact) 设置：与 χ^2 检验设置对话框一致。

选项 (Options) 设置：与 χ^2 检验设置对话框一致。

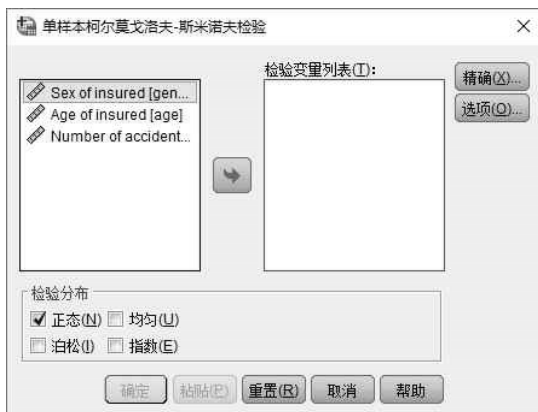




图 8-20 “游程 (Runs test) 参数设置”对话框

8.5.2 实例分析

本实例使用的数据是 SPSS 自带的数据集 autoaccidents.sav，此数据集是机动车事故的调查表，有 gender (性别)、age (年龄)，以及 accident (事故) 三个变量，数据格式如图 8-21 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|----------|----|----|------|-------------------|--------------|----|---|----|----|----|
| 1 | gender | 数字 | 2 | 0 | Sex of insured | {1, Male}... | 无 | 8 | 右 | 标度 | 输入 |
| 2 | age | 数字 | 2 | 0 | Age of insured | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 3 | accident | 数字 | 2 | 0 | Number of acci... | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 8-21 数据集 autoaccidents.sav 的数据格式

-  **结果文件** —— 附带光盘 “PROGRAM\CH08\实例 8-4” 文件夹
-  **动画演示** —— 附带光盘 “AVI\实例 8-4.avi” 文件

1. 参数设置

选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框 单样本 K-S 检验 (1 Sample K-S)”，则弹出如图 8-22 所示对话框，选择变量 Number of accidents past 5 years 到“检验变量列表 (Test Variable List)”选项栏，去掉“常规 (Normal)”选项，选中“泊松 (Poisson)”选项。

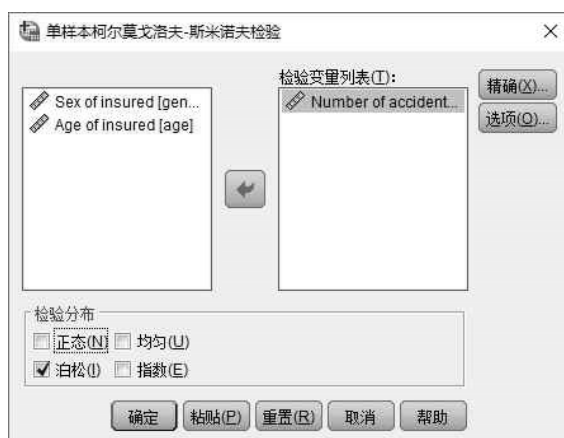


图 8-22 “单样本 K-S 检验参数设置”对话框

2. 结果分析

设置好以后单击主界面的“确定”按钮进行分析。则检验结果如图 8-23 所示。观测样本为 500 个，泊松（Poisson）分布的均值为 1.72，说明在过去的 5 年中平均有 1.72 次事故发生。K-S 样本统计量为 1.460，假设检验的 P 值为 0.028 小于 0.05，所以，不可以确认事故发生次数服从 Poisson 分布，故拒绝原假设。

| 单样本柯尔莫戈洛夫-斯米诺夫检验 | | |
|---------------------|-----|----------------------------------|
| | | Number of accidents past 5 years |
| 个案数 | | 500 |
| 泊松参数 ^{a,b} | 平均值 | 1.72 |
| 最极端差值 | 绝对 | .065 |
| | 正 | .065 |
| | 负 | -.041 |
| 柯尔莫戈洛夫-斯米诺夫 Z | | 1.460 |
| 渐近显著性(双尾) | | .028 |
| a. 检验分布为泊松分布。 | | |
| b. 根据数据计算。 | | |

图 8-23 K-S 检验统计量

下面分组来进行分析，选择菜单“数据（Data） 拆分文件（Split File）”，弹出如图 8-24 所示对话框，选择“按组组织输出”选项栏，并选择变量 Sex of insured 到“分组方式（Groups Based on）”选项栏中，然后单击“确定”按钮。

然后重新打开单样本 K-S 检验对话框，单击“选项（Options）”按钮，弹出如图 8-25 所示对话框，选择“描述性（Descriptive）”选项，然后单击“继续（Continue）”按钮返回主界面。

最后单击主界面中的“确定”按钮进行分组分析。图 8-26 和图 8-27 分别是描述性统计量和 K-S 检验结果。描述性统计量有分组均值、标准差等信息。K-S 检验结果 Male 组的 KOS 统计量为

0.75, 对应的 P 值为 0.627 远大于 0.05, 所以接受原假设, 分布符合 Poisson 分布, 同样对于 Female 组, K-S 统计量为 1.164, P 值为 0.133, 接受原假设, 分布也符合 Poisson 分布。



图 8-24 “拆分文件 Split File 设置”对话框



图 8-25 “选项 (Options) 设置”对话框

| 描述统计 ^a | | | | | |
|----------------------------------|-----|------|-------|-----|-----|
| | 个案数 | 平均值 | 标准差 | 最小值 | 最大值 |
| Number of accidents past 5 years | 250 | 1.98 | 1.608 | 0 | 7 |
| a. Sex of insured = Male | | | | | |

| 描述统计 ^a | | | | | |
|----------------------------------|-----|------|-------|-----|-----|
| | 个案数 | 平均值 | 标准差 | 最小值 | 最大值 |
| Number of accidents past 5 years | 250 | 1.47 | 1.412 | 0 | 6 |
| a. Sex of insured = Female | | | | | |

图 8-26 描述性统计量

| 单样本柯尔莫戈洛夫-斯米诺夫检验 ^a | | | 单样本柯尔莫戈洛夫-斯米诺夫检验 ^a | | |
|-------------------------------|-----|-------|----------------------------------|-----|-------|
| | | | | | |
| | | | Number of accidents past 5 years | | |
| 个案数 ^b | | 250 | 个案数 ^b | | 250 |
| 泊松参数 ^{b,c} | 平均值 | 1.98 | 泊松参数 ^{b,c} | 平均值 | 1.47 |
| 最极端差值 | 绝对 | .047 | 最极端差值 | 绝对 | .074 |
| | 正 | .047 | | 正 | .074 |
| | 负 | -.033 | | 负 | -.042 |
| 柯尔莫戈洛夫-斯米诺夫 Z | | .750 | 柯尔莫戈洛夫-斯米诺夫 Z | | 1.164 |
| 渐近显著性 (双尾) | | .627 | 渐近显著性 (双尾) | | .133 |
| a. Sex of insured = Male | | | a. Sex of insured = Female | | |
| b. 检验分布为泊松分布。 | | | b. 检验分布为泊松分布。 | | |
| c. 根据数据计算。 | | | c. 根据数据计算。 | | |

图 8-27 K-S 统计量

8.6 两个独立样本分布位置检验

如果两个无联系总体的分布是未知的, 则检验两个总体的均值或分布是否有显著差异的方法是一种非参数检验方法, 或者称为两个独立样本的检验。检验是通过两个总体中分别抽取的随机样本数据进行的。

8.6.1 两个独立样本分布位置检验的参数设置

选择菜单“分析（Analyze）→非参数检验（Nonparametric Tests）→2 个独立样本（2 Independent Sample）”，则弹出如图 8-28 所示对话框，对话框主要组成部分如下所述。



图 8-28 “两独立样本（2 Independent Sample）分布位置检验参数设置”对话框

变量选择：如图 8-28 中左上角为待变量列表。

- 检验变量列表（Test Variable List）：检验变量选项框，用于放置检验变量，可以选入多个检验变量。
- 分组变量（Grouping Variable）：分组变量框，用于设置分组变量，选入分组变量后，则激活其下的“定义组（Define Groups）”按钮，单击此按钮则弹出如图 8-29 所示对话框，可以在其中输入两个分组的取值。
- 检验类型（Test Type）：选择检验方法。包括四种方法，曼-惠特尼 U（Mann-Whitney 的 U 检验方法）、莫斯极端反应检验法（Moses extreme reactions）、柯尔莫戈洛夫—斯米诺夫 Z 检验法（Kolmogorov-Smirnov Z）、瓦尔德—沃尔福威茨游程检验法（Wald-Wolfowitz runs）。

精确（Exact）设置：与 χ^2 检验设置对话框一致。

选项（Options）设置：与 χ^2 检验设置对话框一致。



图 8-29 “定义组（Define Groups）设置”对话框

8.6.2 实例分析


本实例所用数据集为 SPSS 自带的 adl.sav 数据集，此数据集为两组调查数据，分别含有 14 个变量，包括 Travel、Co 确定 ing、sm 确定 er 等变量，下面对这两组数据进行分析来确定这两组的能力是否存在显著性的差异，数据集的格式如图 8-30 所示。



| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|----------|----|----|------|---------------------|-----------------|----|---|----|----|----|
| 1 | id | 数字 | 2 | 0 | Pt. ID | 无 | 无 | 8 | 右 | 名义 | 输入 |
| 2 | group | 数字 | 2 | 0 | Treatment group | {0, Control}... | 无 | 8 | 右 | 标度 | 输入 |
| 3 | gender | 数字 | 2 | 0 | Female pts. | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 4 | age | 数字 | 2 | 0 | Pt. age | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 5 | los | 数字 | 2 | 0 | Hospital LOS | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 6 | diabetic | 数字 | 2 | 0 | Diabetes mellitus | {0, No}... | 无 | 8 | 右 | 名义 | 输入 |
| 7 | hypertns | 数字 | 2 | 0 | Hypertensive | {0, No}... | 无 | 8 | 右 | 名义 | 输入 |
| 8 | afib | 数字 | 2 | 0 | Atrial fibrillation | {0, No}... | 无 | 8 | 右 | 名义 | 输入 |
| 9 | priorstr | 数字 | 2 | 0 | Prior stroke | {0, No}... | 无 | 8 | 右 | 名义 | 输入 |

图 8-30 数据集的格式

 **结果文件** —— 附带光盘 “PROGRAM\CH08\实例 8-5” 文件夹

 **动画演示** —— 附带光盘 “AVI\实例 8-5.avi” 文件

首先建立假设：

H_0 ：两组的能力没有显著差异；

H_1 ：两组的能力有显著差异。

1. 参数设置

选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框 两个独立样本 (2 Independent Sample)”，则弹出如图 8-31 所示对话框，选择变量 Travel ADL、CoOKing ADL，以及 Housekeeping ADL 到“检验变量列表 (Test Variable List)”选项栏中，选择变量 Treatment group 到“分组变量 (Grouping Variable)”选项栏中。

然后单击图 8-31 中的“定义组 (Define Groups)”按钮，弹出如图 8-32 所示对话框，在“组 1 (group 1)”和“组 2 (group 2)”选项框中分别填写 0 和 1，然后单击“继续”按钮返回主界面。

2. 结果分析

设置完成以后，单击主界面 2 独立样本检验中的“确定”按钮进行分析，结果如下，首先是秩和表，如图 8-33 所示。

然后是检验统计量的结果，如图 8-34 所示。最后统计量对应的 P 值分别为 0.087、0.03、0.004，其中 $0.087 > 0.05$ ，说明接受原假设 0.03 和 0.004 全小于 0.05，所以拒绝原假设。可以得到在 Travel ADL 上认为两组并无明显的差异，在 Co 确定 ing ADL 和 Housekeepin ADL 上具有显著的差异。



图 8-31 “两独立样本参数设置 (2 Independent Sample)”对话框



图 8-32 “定义组 (Define Groups)”对话框

| 秩 | | | | |
|------------------|-----------------|-----|-------|---------|
| | Treatment group | 个案数 | 秩平均值 | 秩的总和 |
| Travel ADL | Control | 46 | 55.67 | 2561.00 |
| | Treatment | 54 | 46.09 | 2489.00 |
| | 总计 | 100 | | |
| Cooking ADL | Control | 46 | 57.07 | 2625.00 |
| | Treatment | 54 | 44.91 | 2425.00 |
| | 总计 | 100 | | |
| Housekeeping ADL | Control | 46 | 59.30 | 2728.00 |
| | Treatment | 54 | 43.00 | 2322.00 |
| | 总计 | 100 | | |

图 8-33 秩和表

| 检验统计 ^a | | | |
|-------------------|------------|-------------|----------------------|
| | Travel ADL | Cooking ADL | Housekeepin g ADL |
| 曼-惠特尼 U | 1004.000 | 940.000 | 837.000 |
| 威尔科克森 W | 2489.000 | 2425.000 | 2322.000 |
| Z | -1.711 | -2.165 | -2.914 |
| 渐近显著性 (双尾) | .087 | .030 | .004 |

a. 分组变量: Treatment group

图 8-34 检验统计量结果

8.7 多个独立样本分布位置检验

在总体分布未知的情况下，多个独立样本的检验是检验多个独立总体的平均值是否存在显著的差异。由于总体分布未知，所以检验过程是在建立秩的基础上。下面通过例题来说明具体的检验方法，首先是 SPSS 的参数设置。

8.7.1 多个独立样本分布位置检验的参数设置

选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框 K 个独立样本 (K Independent Sample)”，则弹出如图 8-35 所示对话框，对话框主要组成部分

如下所述。

变量选择：如图 8-35 中左上角为待变量列表。

- 检验变量列表 (Test Variable List)：检验变量选项框，用于放置检验变量，可以选入多个检验变量。
- 分组变量 (Grouping Variable)：分组变量框，用于设置分组变量，选入分组变量后，则激活其下的“定义范围 (Define Range)”按钮，单击此按钮则弹出如图 8-36 所示对话框，可以在其中输入最小和最大取值，取值在外的观测值将被排除在检验分析之外。
- 检验类型 (Test Type)：选择检验方法。包括三种方法，克鲁斯卡尔—沃利斯检验方法 (Kruskal-Wallis H)、中位数检验方法 (Median)、约克海—塔帕斯特拉检验法 (Jonckheere-Terpstra)。

精确 (Exact) 设置：与 χ^2 检验设置对话框一致。

选项 (Options) 设置：与 χ^2 检验设置对话框一致。



图 8-35 “多个独立样本 (K Independent Sample) 分布位置检验参数设置”对话框



图 8-36 “定义范围 (Define Range) 设置”对话框

8.7.2 实例分析

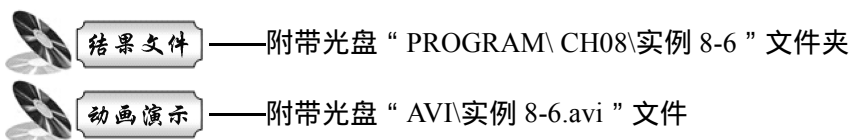
本实例所用数据为 SPSS 自带的数据集 salesperformance.sav，数据集为三组销售绩效数据，下面就对这三组数据进行分析，来考察三组的销售绩效是否存在显著差异？数据格式如图 8-37 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|---------|----|----|------|---------------------|---|----|---|----|----|----|
| 1 | group | 数字 | 1 | 0 | Sales training g... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 2 | perform | 数字 | 8 | 2 | Score on trainin... | 无 | 无 | 8 | 右 | 标度 | 输入 |

数据视图 变量视图

IBM SPSS Statistics 处理程序就绪 Unicode ON

图 8-37 数据集 salesperformance.sav 的数据格式



1. 参数设置

选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框 K 个独立样本 (K Independent Sample)”, 则弹出如图 8-38 所示对话框, 选择变量 Score on training exam 到“检验变量列表 (Test Variable List)”选项栏中, 去掉克鲁斯卡尔—沃利斯检验选项, 选择中位数选项, 选择变量 Sales training group 到“分组变量”选项栏。



图 8-38 “多个独立样本 (K Independent Sample) 设置”对话框

然后单击“定义范围 (Define Range)”按钮, 弹出如图 8-39 所示对话框, 在“最小值”和“最大值”选项栏中分别填入 1 和 3, 然后单击“继续 (Continue)”按钮返回主界面。

然后单击图 8-38 中的“选项 (Options)”按钮, 弹出如图 8-40 所示对话框, 选择“四分位数 (Quartiles)”选项栏, 然后单击“继续 (Continue)”按钮返回主界面, 最后再单击主界面检验类型中的中位数。



图 8-39 “定义范围 (Define Range) 设置”对话框 图 8-40 “选项 (Options) 设置”对话框

2. 结果分析

设置完成以后, 单击主界面“K 个独立样本设置”对话框中的“确定”按钮进行分析, 结果如图 8-41 所示, 首先是描述性的统计量, 包括观测数、中位数。

然后是频率表，如图 8-42 所示，分组给出频率。

| 描述统计 | | | | |
|------------------------|-----|---------|--------------|---------|
| | 个案数 | 百分位数 | | |
| | | 第 25 个 | 第 50 个 (中位数) | 第 75 个 |
| Score on training exam | 60 | 64.6322 | 74.9330 | 81.3159 |
| Sales training group | 60 | 1.00 | 2.00 | 3.00 |

图 8-41 描述性统计量

| 频率 | | | | |
|------------------------|-------|----------------------|----|----|
| | | Sales training group | | |
| | | 1 | 2 | 3 |
| Score on training exam | > 中位数 | 4 | 11 | 15 |
| | ≤ 中位数 | 16 | 9 | 5 |

图 8-42 频率表

最后是检验统计量，如图 8-43 所示，可以得到卡方统计量为 12.400，其对应的 P 值为 0.002 小于 0.05，所以拒绝原假设，认为这三组的成绩具有显著的差异性。

8.8 两个相关样本分布位置检验

两个相关样本检验一般用于比较一个现象在采取了某项措施前后的变化是否显著，或者说采取的措施是否有效。也可以检验同一个测试对象上的两种测试方法是否一致。取 n 个测试对象作为样本，则样本数据是成对出现的。也可以检验这样两个样本是否服从相同的分布等。这种检验在实际中应用范围很广，如对于一种药品效果比较检验，农业上对于一种新的粮食品种与原有品种的比较检验，工业中新工艺方法、新材料与原方法和材料的比较检验等。

| 检验统计 ^a | |
|------------------------|---------------------|
| Score on training exam | |
| 个案数 | 60 |
| 中位数 | 74.9330 |
| 卡方 | 12.400 ^b |
| 自由度 | 2 |
| 渐近显著性 | .002 |

a. 分组变量: Sales training group
b. 0 个单元格 (0.0%) 的期望频率低于 5。期望的最低单元格频率为 10.0。

图 8-43 检验统计量

8.8.1 两个相关样本分布位置检验的参数设置

选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框 2 个相关样本 (2 Relate Sample)”，则弹出如图 8-44 所示对话框，对话框主要组成部分如下所述。



图 8-44 “2 个相关样本 (2 Relate Sample) 分布位置检验参数设置”对话框

变量选择：如图 8-44 中左上角为待变量列表。

- 检验对 (Test Pairs): 用于选入相关检验变量列表框。
- 检验类型 (Test Type): 选择检验方法。包括四种方法, 威尔科克森检验法 (Wilcoxon)、符号检验法 (Sign)、麦克尼马尔检验法 (McNemar)、边际齐性法 (Marginal Homogeneity)。
精确 (Exact) 设置: 与 χ^2 检验设置对话框一致。
选项 (Options) 设置: 与 χ^2 检验设置对话框一致。

8.8.2 实例分析

一车间为了提高工作效率, 对某种零件的加工过程进行改进, 为了比较加工时间是否明显减少, 抽取 15 名工人对比他们改革前后零件的加工时间, 得到相应的数据如下: 试根据数据检验改进后零件的加工时间是否明显减少?

改进前 (m): 70, 76, 56, 63, 63, 56, 58, 60, 65, 65, 75, 66, 56, 59, 70

改进后 (m): 48, 54, 60, 64, 48, 55, 54, 45, 51, 48, 56, 48, 64, 50, 54



结果文件

——附带光盘 “PROGRAM\CH08\实例 8-7” 文件夹



动画演示

——附带光盘 “AVI\实例 8-7.avi” 文件

首先建立假设:

H_0 : 改进前后的零件加工时间没有显著差异。

H_1 : 改进前后的零件加工时间明显减少。

1. 参数设置

单击 “Analyze → Nonparametric Test 旧对话框 → 两个相关样本 (2 Relate Sample)”, 打开 “2 个相关样本 (2 Relate Sample)” 对话框如图 8-45 所示。选中变量 first 和 second 到 “检验对 (Test Pair(s))” 选项框中, 在 “检验类型 (Test Type)” 选项栏中选择检验方式。SPSS 中给出了几种检验方法, 本例中选择威尔科克森和符号检验法, 如图 8-45 所示。



图 8-45 “2 个相关样本 (2 Relate Sample) 检验设置” 对话框

然后单击“选项”按钮，弹出如图 8-46 所示对话框，选中“描述性”选项栏，然后单击“继续”按钮返回主界面。

2. 结果分析

设置完成以后单击主界面的“确定”按钮进行分析，结果如下。图 8-47 为威尔科克森检验结果，图 8-48 是威尔科克森检验统计量，图 8-49 是符号检验频率表，图 8-50 是符号检验统计量。

威尔科克森检验结果检验统计量 Z 值为 -2.842 ，假设检验的 P 值为 0.004 小于 0.05 。符号检验统计量的 P 值为 0.035 也小于 0.05 ，所以拒绝原假设，认为改进前后的差异是显著的。



图 8-46 “选项 (Options) 设置”对话框

| 秩 | | | | |
|-----------|-----|-----------------|------|--------|
| | | 个案数 | 秩平均值 | 秩的总和 |
| 改进后 - 改进前 | 负秩 | 12 ^a | 9.17 | 110.00 |
| | 正秩 | 3 ^b | 3.33 | 10.00 |
| | 绑定值 | 0 ^c | | |
| 总计 | | 15 | | |

a. 改进后 < 改进前
b. 改进后 > 改进前
c. 改进后 = 改进前

图 8-47 威尔科克森检验结果

| 检验统计 ^a | |
|-------------------|---------------------|
| | 改进后 - 改进前 |
| Z | -2.842 ^b |
| 渐近显著性 (双尾) | .004 |

a. 威尔科克森符号秩检验
b. 基于正秩。

图 8-48 威尔科克森检验统计量

| 频率 | | |
|-----------|------------------|-----|
| | | 个案数 |
| 改进后 - 改进前 | 负差值 ^a | 12 |
| | 正差值 ^b | 3 |
| | 绑定值 ^c | 0 |
| 总计 | | 15 |

a. 改进后 < 改进前
b. 改进后 > 改进前
c. 改进后 = 改进前

图 8-49 符号检验频率表

| 检验统计 ^a | |
|-------------------|-------------------|
| | 改进后 - 改进前 |
| 精确显著性 (双尾) | .035 ^b |

a. 符号检验
b. 使用了二项分布。

图 8-50 符号检验统计量

8.9 多个相关样本分布位置检验

多个有联系样本的方差分析，又称多个相关样本的检验，是在总体分布未知的情况下，用于比较多个有联系的总体分布的差异性，可以归纳如下。

- 多个有联系的总体是否存在显著差异。
- 多个评判结果是否存在显著差异（一致性检验）。

由于总体分布未知，所以检验都是建立在秩和的基础上。

8.9.1 多个相关样本分布位置检验的参数设置

选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Tests) 旧对话框 K 个相关样本 (K-Related Samples)”，则弹出如图 8-51 所示对话框，对话框主要组成部分如下所述。

变量选择：如图 8-51 中左侧为待变量列表。

- 检验对 (Test Pairs)：用于选入相关检验变量列表框。
- 检验类型 (Test Type)：选择检验方法。包括三种方法，傅莱德曼 (Friedman) 检验、肯德尔 W (Kendall) 检验、柯克兰 Q (Cochran) 检验。

精确 (Exact) 设置：与 χ^2 检验设置对话框一致。

选项 (Options) 设置：与 χ^2 检验设置对话框一致。

8.9.2 实例分析



本实例使用 SPSS 自带的数据集 webusability.sav，此数据集为 6 个任务的实施情况，其中 0 代表失败，1 代表成功，下面就对这 6 个任务是否具有差异性进行分析。数据集 webusability.sav 如图 8-52 所示。



图 8-51 “多个相关样本 (K-Related Samples) 分布位置检验参数设置”对话框

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|-------|----|----|------|--------------------|-----------------|----|---|----|----|----|
| 1 | task1 | 数字 | 1 | 0 | Registered warr... | {0, Failure}... | 无 | 8 | 右 | 名义 | 输入 |
| 2 | task2 | 数字 | 1 | 0 | Received auto... | {0, Failure}... | 无 | 8 | 右 | 名义 | 输入 |
| 3 | task3 | 数字 | 1 | 0 | Received introd... | {0, Failure}... | 无 | 8 | 右 | 名义 | 输入 |
| 4 | task4 | 数字 | 1 | 0 | Added question... | {0, Failure}... | 无 | 8 | 右 | 名义 | 输入 |
| 5 | task5 | 数字 | 1 | 0 | Updated shoppi... | {0, Failure}... | 无 | 8 | 右 | 名义 | 输入 |
| 6 | task6 | 数字 | 1 | 0 | Edited databas... | {0, Failure}... | 无 | 8 | 右 | 名义 | 输入 |

图 8-52 数据集 webusability.sav

-  **结果文件** —— 附带光盘 “PROGRAM\CH08\实例 8-8” 文件夹
-  **动画演示** —— 附带光盘 “AVI\实例 8-8.avi” 文件

1. 参数设置

在数据窗口中选择菜单“分析 (Analyze) 非参数检验 (Nonparametric Test) 旧对话框 K 个相关样本 (K Relate Sample)”，打开“K 个相关样本 (K Relate Sample)”对话框如图 8-53 所示。选中所有变量到“检验变量 (Test Variables)”选项栏中，然后去掉“布莱德曼”选项，选中“柯克兰 Q”的选项栏。



图 8-53 “K 个相关样本 (K Relate Sample)”对话框

然后单击图 8-53 中的“统计量 (Statistics)”按钮，弹出如图 8-54 所示对话框，选中“描述性 (Descriptive)”选项，并单击“继续”按钮返回主界面。

2. 结果分析

设置完成以后单击主界面“K 个相关样本检验”对话框的“确定”按钮进行分析，结果如下，首先是如图 8-55 所示的描述性统计量，包括均值、标准差、最大值、最小值等信息。



图 8-54 “统计量设置”对话框

| 描述统计 | | | | | |
|------------------------------------|-----|------|------|-----|-----|
| | 个案数 | 平均值 | 标准差 | 最小值 | 最大值 |
| Registered warranty data | 5 | 1.00 | .000 | 1 | 1 |
| Received automated fax information | 5 | .40 | .548 | 0 | 1 |
| Received introductory newsletter | 5 | .40 | .548 | 0 | 1 |
| Added question to support list | 5 | .00 | .000 | 0 | 0 |
| Updated shopping cart | 5 | .80 | .447 | 0 | 1 |
| Edited database information | 5 | .80 | .447 | 0 | 1 |

图 8-55 描述性统计量

然后是频率表,如图 8-56 所示。

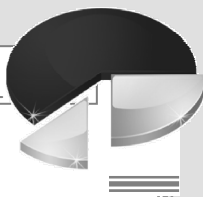
最后是检验统计量,如图 8-57 所示。Cochran 的 Q 统计量为 12.949,对应的 P 值为 0.024 小于 0.05,所以拒绝原假设,即 6 个任务并不具有显著的差异性。

| 频率 | 值 | |
|------------------------------------|---|---|
| | 0 | 1 |
| Registered warranty data | 0 | 5 |
| Received automated fax information | 3 | 2 |
| Received introductory newsletter | 3 | 2 |
| Added question to support list | 5 | 0 |
| Updated shopping cart | 1 | 4 |
| Edited database information | 1 | 4 |

图 8-56 频率表

| 检验统计 | |
|-------------|---------------------|
| 个案数 | 5 |
| 柯克兰 Q | 12.949 ^a |
| 自由度 | 5 |
| 渐近显著性 | .024 |
| a. 1 被视为成功。 | |

图 8-57 检验统计量



第 9 章 方差分析

方差分析 (ANOVA) 又称“变异数分析”或“F 检验”, 是 R.A.Fisher 发明的, 用于两个及两个以上样本均数差别的显著性检验。基本思想是通过分析研究中不同来源的变异对总变异的贡献大小, 从而确定可控因素对研究结果影响力的大小。主要应用于均数差别的显著性检验; 分离各有关因素并估计其对总变异的作用; 分析因素间的交互作用; 方差齐性检验。



本讲内容

- 方差分析的基本原理
- 单因素 ANOVA 检验
- 多因素方差分析
- 协方差分析

9.1 方差分析的基本原理

方差分析由于各种因素的影响, 研究所得的数据呈现波动状。造成波动的原因可分成两类, 一是不可控的随机因素; 二是研究中施加的对结果形成影响的可控因素。

一个复杂的事物, 其中往往有许多因素互相制约又互相依存。方差分析的目的是通过数据分析找出对该事物有显著影响的因素, 各因素之间的交互作用, 以及显著影响因素的最佳水平等。方差分析是在可比较的数组中, 把数据间的总的“变差”按各指定的变差来源进行分解的一种技术。对变差的度量, 采用离差平方和。方差分析方法就是从总离差平方和分解出可追溯到指定来源的部分离差平方和, 这是一个很重要的思想。

经过方差分析若拒绝了检验假设, 只能说明多个样本总体均数不相等或不全相等。若要得到各组均数间更详细的信息, 应在方差分析的基础上进行多个样本均数的事后比较。

在讨论方差分析之前, 先了解一些常用的术语。

① 试验指标 (Experimental Index): 为衡量试验结果的好坏或处理效应的高低, 在试验中具体测定的性状或观测的项目称为试验指标。由于试验目的不同, 选择的试验指标也

不相同。在畜禽、水产试验中常用的试验指标有日增重、产仔数、产奶量、产蛋率、瘦肉率、某些生理生化指标和体型指标（如血糖含量、体高、体重）等。

② 试验因素 (Experimental 因子 (Factor))：试验中所研究的影响试验指标的因素叫试验因素。如研究如何提高猪的日增重时，饲料的配方、猪的品种、饲养方式、环境温度等都对日增重有影响，均可作为试验因素来考虑。当试验中考察的因素只有一个时，称为单因素试验；若同时研究两个或两个以上的因素对试验指标的影响时，则称为两因素或多因素试验。试验因素常用大写字母 A 、 B 、 C 、... 等表示。

③ 因素水平 (Level of 因子 (Factor))：试验因素所处的某种特定状态或数量等级称为因素水平，简称水平。如比较 3 个品种奶牛产奶量的高低，这 3 个品种就是奶牛品种这个试验因素的 3 个水平；研究某种饲料中 4 种不同能量水平对肥育猪瘦肉率的影响，这 4 种特定的能量水平就是饲料能量这一试验因素的 4 个水平。因素水平用代表该因素的字母加添足标 $1, 2, \dots$ ，来表示。如 A_1 、 A_2 、...， B_1 、 B_2 、...，等。

④ 试验处理 (Treatment)：事先设计好的实施在试验单位上的具体项目叫试验处理，简称处理。在单因素试验中，实施在试验单位上的具体项目就是试验因素的某一水平。例如，进行饲料的比较试验时，实施在试验单位（某种畜禽）上的具体项目就是喂饲某一种饲料。所以进行单因素试验时，试验因素的一个水平就是一个处理。在多因素试验中，实施在试验单位上的具体项目是各因素的某一水平组合。例如，进行 3 种饲料和 3 个品种对猪日增重影响的两因素试验，整个试验共有 $3 \times 3 = 9$ 个水平组合，实施在试验单位（试验猪）上的具体项目就是某品种与某种饲料的结合。所以，在多因素试验时，试验因素的一个水平组合就是一个处理。

⑤ 试验单位 (Experimental Unit)：在试验中能接受不同试验处理的独立的试验载体称为试验单位。在畜禽、水产试验中，一只家禽、一头家畜、一只小白鼠、一尾鱼，即一个动物；或几只家禽、几头家畜、几只小白鼠、几尾鱼，即一组动物都可作为试验单位。试验单位往往也是观测数据的单位。

⑥ 重复 (Repetition)：在试验中，将一个处理实施在两个或两个以上的试验单位上，称为处理有重复；一处理实施的试验单位数称为处理的重复数。例如，用某种饲料喂 4 头猪，就说这个处理（饲料）有 4 次重复。

9.1.1 自由度与平方和分解

方差是平方和除以自由度的商。要将一个试验资料的总变异分解为各个变异来源的相应变异，首先将总平方和与总自由度分解为各个变异来源的相应部分。因此，平方和与自由度的分解是方差分析的第一步。下面以单因素完全随机试验设计的资料为例说起。

假设有 k 个处理，每个处理有 n 个观察值，则该试验资料共有 nk 个观察值，其观察值的组成参见表 9-1。表 9-1 中， i 代表资料中任一样本； j 代表样本中任一观测值； x_{ij} 代表任一样本的任一观测值； T_i 代表处理总和； \bar{x}_i 代表处理平均数； T 代表全部观测值总和； \bar{x} 代表总平均数。

表 9-1 每处理具 n 个观测值的 k 组数据的符号表

| 处 理 | 观 察 值 | | | | | | 处理总和 T_i | 处理平均 \bar{x}_i |
|-----|----------|----------|-----|----------|-----|----------|---------------|---------------------|
| | 1 | 2 | ... | J | ... | N | | |
| 1 | x_{11} | x_{12} | ... | x_{1j} | ... | x_{1n} | T_{1i} | \bar{x}_{1i} |
| 2 | x_{21} | x_{22} | ... | x_{2j} | ... | x_{2n} | T_{2i} | \bar{x}_{2i} |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| I | x_{i1} | x_{i2} | ... | x_{ij} | ... | x_{in} | T_{ii} | \bar{x}_{ii} |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| K | x_{k1} | x_{k2} | ... | x_{kj} | ... | x_{kn} | T_{ik} | \bar{x}_{ik} |
| | | | | | | | $T = \sum x$ | \bar{x} |

在表 9-1 中，总变异是 nk 个观测值的变异，故其自由度 $\nu = nk - 1$ ，而其平方和则为

$$SS_T = \sum_{i=1}^{nk} (x_{ij} - \bar{x})^2 = \sum x^2 - C$$

式中， C 称为矫正数，即

$$C = \frac{(\sum x)^2}{nk} = \frac{T^2}{nk}$$

产生总变异的原因可从两方面来分析：一是同一处理不同重复观测值的差异是由偶然因素影响造成的，即试验误差，又称组内变异；二是不同处理之间平均数的差异主要是由处理的不同效应所造成，称处理间变异，又称组间变异。因此，总变异可分解为组间变异和组内变异两部分。

组间的差异即 k 个 \bar{x} 的变异，故自由度 $\nu = k - 1$ ，而其平方和为

$$SS_t = n \sum_{i=1}^k (\bar{x}_{ij} - \bar{x})^2 = \frac{\sum T_i^2}{n} - C$$

组内的变异为各组内观测值与组平均数的变异，故每组具有自由度 $\nu = n - 1$ 和平方和

$\sum_{i=1}^n (x_{ij} - \bar{x})^2$ ，而资料共有 k 组，故组内自由度 $\nu = k(n - 1)$ ，而组内平方和为

$$SS_e = \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 = SS_T - SS_t$$

因此，得到表 9-1 类型资料平方和与自由度的分解式为

总平方和 = 组间（处理间）平方和 + 组内（误差）平方和

$$\sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x})^2 = n \sum_{i=1}^k (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2$$

记为

$$SS_T = SS_t + SS_e$$

总自由度 = 组间（处理间）自由度 + 组内（误差）自由度

即

$$nk-1=(k-1)+k(n-1)$$

记为

$$DF_T=DF_t+DF_e$$

将以上公式归纳为以下几个方面。

总平方和： $SS_T=\sum x^2-C$ ；总自由度： $DF_T=kn-1$ 。

处理平方和： $SS_t=\frac{\sum T_i^2}{n}-C$ ；处理自由度： $DF_t=k-1$ 。

误差平方和： $SS_e=SS_T-SS_t$ ；误差自由度： $DF_e=k(n-1)$ 。

求得各变异来源的平方和与自由度后，进而求得

总的方差： $S_T^2=\frac{SS_T}{DF_T}$ ；处理间方差： $S_t^2=\frac{SS_t}{DF_t}$ ；误差方差： $S_e^2=\frac{SS_e}{DF_e}$ 。

9.1.2 F 检验

1. F 分布

设想在一正态总体 $N(\mu, \sigma^2)$ 中随机抽取样本容量为 n 的样本 k 个，将各样本观测值整理成表 9-1 的形式。此时的各处理没有真实差异，各处理只是随机分的组。因此，由上述公式算出的 S_t^2 和 S_e^2 都是误差方差 σ^2 的估计量。以 S_e^2 为分母， S_t^2 为分子，求其比值。统计学上把两个方差之比值称为 F 值，即

$$F=S_t^2/S_e^2$$

F 分布具有两个自由度： $\nu_1=df_t=k-1, \nu_2=df_e=k(n-1)$ 。 F 值所具有的概率分布称为 F 分布。

F 分布的取值范围为 $(0, +\infty)$ ，其平均值 $\mu_F=1$ 。用 $f(F)$ 表示 F 分布的概率密度函数，则其分布函数 $F(F_\alpha)$ 为

$$F(F_\alpha)=P(F < F_\alpha)=\int_0^{F_\alpha} f(F)dF$$

因而 F 分布右尾从 F_α 到 $+\infty$ 的概率为

$$P(F > F_\alpha)=1-F(F_\alpha)=\int_{F_\alpha}^{+\infty} f(F)dF$$

2. F 检验

F 值表是专门为检验 S_t^2 代表的总体方差是否比 S_e^2 代表的总体方差大而设计的。若实际计算的 F 值大于 $F_{0.05}$ ，则 F 值在 $\alpha=0.05$ 的水平上显著，以 95% 的可靠性（冒 5% 的风险）推断 S_t^2 代表的总体方差大于 S_e^2 代表的总体方差。这种用 F 值出现概率的大小推断两个总体方差是否相等的方法称为 F 检验。

在方差分析中所进行的 F 检验目的在于推断处理间的差异是否存在，检验某项变异因素的效应方差是否为零。因此，在计算 F 值时总是以被测验因素的方差作分子，以误差方

差作分母。应当注意,分母项的正确选择是由方差分析的模型和各项变异原因的期望均方决定的。

实际进行 F 检验时,是将由试验资料所算得的 F 值与根据 $v_2=DF_t$ (大均方,即分子均方的自由度) $v_2=DF_e$ (小均方,即分母均方的自由度) 查附表 F 值表所得的临界 F 值与 $F_{0.05}$ 、 $F_{0.01}$ 相比较做出统计推断的。

若 $F < F_{0.05}$, 即 $P > 0.05$, 不能否定 H_0 , 统计学上把这一检验结果表述为,各处理间差异不显著,不标记符号;若 $F_{0.05} < F < F_{0.01}$, 即 $0.01 < P < 0.05$, 否定 H_0 , 接受 H_A , 统计学上,把这一检验结果表述为,各处理间差异显著,在 F 值的右上方标记“*”;若 $F > F_{0.01}$, 即 $P < 0.01$, 否定 H_0 , 接受 H_A , 统计学上,把这一检验结果表述为,各处理间差异极显著,在 F 值的右上方标记“**”。

在实际进行方差分析时,只需计算出各项平方和与自由度,各项均方的计算及 F 检验可在方差分析表上进行。

9.1.3 多重比较

经 F 检验,差异达到显著或极显著,表明试验的总变异主要来源于处理间的变异,试验中各处理平均数间存在显著或极显著差异,但并不意味着每两个处理平均数间的差异都显著或极显著,也不能具体说明哪些处理平均数间有显著或极显著差异,哪些差异不显著。因而,有必要进行两两处理平均数间的比较,以具体判断两两处理平均数间的差异显著性。统计上把多个平均数两两间的相互比较称为多重比较 (Multiple Comparison)。

多重比较的方法比较多,常用的有最小显著差数法和最小显著极差法,现分别进行一下介绍。

1. 最小显著差数法

最小显著差数法 (Least Significant Difference) 又称 LSD 法。此方法是多重比较中最基本的方法。它是两个平均数相比较在多样本试验中的应用,所以, LSD 法实际上属于 t 检验性质的,而 t 检验只适用于测验两个相互独立的样本平均数的差异显著性。在多个平均数时,任何两个平均数比较会牵连到其他平均数,从而降低了显著水平,容易作出错误的判断。所以,在应用 LSD 法进行多重比较时,必须在检验显著的前提下进行,并且各对被比较的两个样本平均数在试验前已经指定,因而它们是相互独立的。利用此法时,各试验处理一般是与指定的对照相比较。

LSD 法的步骤如下所述。

第一步:先计算样本平均数差数标准误 $s_{\bar{x}_1 - \bar{x}_2}$ 为

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s_e^2}{n}}$$

第二步:计算出显著水平为 α 的最小显著差数 LSD_α 。在 t 检验中已知

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

在误差自由度下,查显著水平为 α 时的临界 t 值,令上式 $t = t_\alpha$, 移项可得

$$\bar{x}_1 - \bar{x}_2 = t_a \times s_{\bar{x}_1 - \bar{x}_2}$$

故 $\bar{x}_1 - \bar{x}_2$ 等于在误差自由度下, 显著水平为 α 时的最小显著差数, 即

$$LSD_a = t_a \times s_{\bar{x}_1 - \bar{x}_2}$$

当 $\alpha=0.05$ 和 0.01 时, LSD 的计算公式分别为

$$LSD_{0.05} = t_{0.05} \times s_{\bar{x}_1 - \bar{x}_2}$$

$$LSD_{0.01} = t_{0.01} \times s_{\bar{x}_1 - \bar{x}_2}$$

任何两处理平均数的差数达到或超过 $LSD_{0.05}$ 时, 差异显著; 达到或超过 $LSD_{0.01}$ 时, 差异达到极显著。

2. 新复极差法

新复极差法又称最小显著极差法 (Shortest Significant Ranges, SSR), 目前在农业科学研究中普遍应用。此法的特点是将平均数按照大小进行排序, 不同的平均数之间比较采用不同的显著标准。可用全距相当于平均数标准误的倍数 (SSR) 来衡量, 即

$$\frac{R}{s_{\bar{x}}} = SSR_{\alpha}$$

式中, R 为全距; $s_{\bar{x}}$ 为样本平均数的标准误为

$$s_{\bar{x}} = \sqrt{\frac{s_e^2}{n}}$$

如果 $\frac{R}{s_{\bar{x}}} \geq SSR_{0.05}$, 说明差异显著; $\frac{R}{s_{\bar{x}}} \geq SSR_{0.01}$, 说明差异极显著。将这两个不等式

转换成以下公式, 即

$$R \geq SSR_{0.05} \times s_{\bar{x}} = LSR_{0.05}, \text{ 差异显著}$$

$$R \geq SSR_{0.01} \times s_{\bar{x}} = LSR_{0.01}, \text{ 差异极显著}$$

公式中的 SSR_{α} 为在 α 水平上的最小显著极差。 SSR_{α} 数值的大小, 一方面与误差方差的自由度有关, 另一方面与测验极差所包括的平均数个数 k 有关。

9.2 单因素 ANOVA 检验

在方差分析中, 根据所研究试验因素的多少, 可分为单因素、两因素和多因素试验资料的方差分析。单因素试验资料的方差分析是最简单的一种, 目的在于正确判断该试验因素各水平的优劣。根据各处理内重复数是否相等, 单因素 ANOVA 检验又分为重复数相等和重复数不等两种情况。

在单因素 ANOVA 检验中, 若因素 A 共有 r 个水平, 对均衡试验而言, 每个水平的样本容量为 k , 则共有 kr 个观察值, 参见表 9-2。对不均衡试验, 各水平中的样本容量可以是不相同的, 设这个变量是 n_i 。

表 9-2 单因素 ANOVA 检验的数据结构

| 观测值 j 水平 i | | 1 | 2 | ... | k |
|-------------------|----------|----------|----------|----------|----------|
| 因素 A | 水平 1 | x_{11} | x_{12} | ... | x_{1k} |
| | 水平 2 | x_{21} | x_{22} | ... | x_{2k} |
| | \vdots | \vdots | \vdots | \vdots | \vdots |
| | 水平 r | x_{r1} | x_{r2} | ... | x_{rk} |

9.2.1 单因素 ANOVA 检验步骤

标准的单因素 ANOVA 检验模型为

$$x_{ij} = \mu + a_i + \varepsilon_{ij}$$

式中, x_{ij} 表示第 i 组的第 j 个观察值; μ 表示总体的平均水平; a_i 表示影响因素在 i 水平下对应变量的附加效应, 所有 a_i 之和应当为 0; ε_{ij} 为一个服从正态分布 $N(0, \sigma^2)$ 的随机变量, 代表随机误差。

一般情况下, 做假设检验实际上就是检验各个 a_i 是否均为 0, 如都为 0, 即各组总体均值都相等, 则 x_{ij} 称为服从正态分布的一个变量。

为便于理解, 可以简化模型, 即

$$x_{ij} = \mu_i + \varepsilon_{ij}$$

此时检验的含义变成了各组均值 μ_i 是否相同, 因此原假设设定为 $H_0: \mu_1 = \mu_2 = \cdots = \mu_r$; 而备择假设为 $H_1: \mu_1, \mu_2, \cdots, \mu_r$ 不全等。

构造检验 F 统计量

(1) 水平的均值

令 \bar{x}_i 为第 i 种水平的样本均值, 则

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$$

当各水平的观察值个数均相等的时候, 公式变为

$$\bar{x}_i = \frac{\sum_{j=1}^k x_{ij}}{k}$$

(2) 全部观察值的总均值

令 $\bar{\bar{x}}$ 为全部观察值的总均值, 则

$$\bar{\bar{x}} = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} x_{ij}}{rn_i}$$

当各水平的观察值个数均相等的时候, 公式变为

$$\bar{\bar{x}} = \frac{\sum_{i=1}^r \sum_{j=1}^k x_{ij}}{rk} = \frac{\sum_{i=1}^r \bar{x}_i}{r}$$

(3) 离差平方和

在单因素 ANOVA 检验中, 离差平方和有如下三个。

总离差平方和 (Sum of Squares for Total, SST), 计算公式为

$$SST = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2$$

总离差平方和反映全部观察值的离散状况, 是全部观察值与总平均值的离差平方和。

误差项离差平方和 (Sum of Squares for Error, SSE), 计算公式为

$$SSE = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

误差项离差平方和又称组内离差平方和, 它反映了水平内部观察值的离散情况, 即随机因素产生的影响。

水平项离差平方和 (Sum of Squares for Factor A, SSA)。计算公式为

$$SSA = \sum_{i=1}^r n_i (\bar{x}_i - \bar{\bar{x}})^2$$

水平项离差平方和又称组间离差平方和, 是各组平均值与总平均值的离差平方和。它既包括随机误差, 也包括系统误差。

由于各样本的独立性, 使得变差具有可分解性, 即总离差平方和等于误差项离差平方和加上水平项离差平方和, 用公式表达为

$$SST = SSE + SSA$$

均方, 各离差平方和的大小与观察值的多少有关, 为了消除观察值多少对离差平方和大小的影响, 需要将其平均, 这就是均方。计算方法是用离差平方和除以相应的自由度 df 。

构造检验统计量为

$$F = \text{组间方差/组内方差} = MSA/MSE$$

9.2.2 判断与结论

在假设条件成立时, F 统计量服从第一自由度 df_1 为 $r-1$ 、第二自由度 df_2 为 $n-r$ 的 F 分布。将统计量 F 与给定的显著性水平 α 的临界值 $F_\alpha(r-1, n-r)$ 比较, 可以做出拒绝或不能拒绝原假设 H_0 的决策。

若 $F \geq F_\alpha$, 则拒绝原假设 H_0 , 表明均值之间的差异显著, 因素 A 对观察值有显著影响;

若 $F < F_\alpha$, 则不能拒绝原假设 H_0 , 表明均值之间的差异不显著, 因素 A 对观察值没有显著影响。

9.2.3 单因素 ANOVA 检验过程的参数设置

选择菜单“分析 (Analyze) 比较平均值 (Compare Means) 单因素 ANOVA 检验 (One-Way ANOVA)”，则弹出如图 9-1 所示的对话框，此对话框主要有以下几部分组成。

1. 变量选择设置

图 9-1 所示的左边是变量选项框。

因变量列表 (Dependent List) 选项栏：选择单因素方差分析的指标变量，可以同时选择多个变量，系统会分别对各个指标做单因素方差分析。

因子 (Factor) 选项栏：用于选择因素变量，只可以选择一个变量。

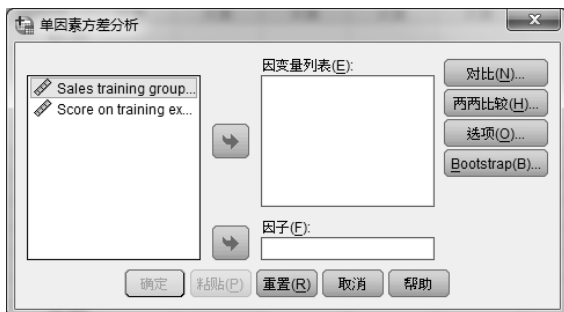


图 9-1 “单因素 ANOVA 检验 (One-Way ANOVA) 参数设置”对话框

2. 对比 (Contrasts) 设置

选择“对比 (Contrasts)”按钮，则弹出如图 9-2 所示对话框，此对话框用于对组间平方和进行分解并确定均值的多项式比较。

- 多项式 (Polynomial)：选择是否对方差分析的组间平方和进行分解并进行趋势检验。
- 等级 (Degree) 下拉菜单：选中多项式后，此下拉菜单被激活，用于选择进行趋势检验的曲线类型。
- 第 1/1 项对比 (Contrast 1 of 1)：精确定义均值比较的多项式系数。添加、更改、删除按钮分别用于添加、修改和删除。

3. 事后比较 (Post Hoc) 设置

选择“事后比较 (Post Hoc)”按钮，则弹出如图 9-3 所示对话框，此对话框用于定义多重比较的检验方法。

(1) 假定等方差 (Equal Variances Assumed) 选项栏

此栏用于定义当样本方差齐次情况下多重比较的检验方法，共有 14 种方法，其中 LSD-N-K 为最常用的方法。

- LSD：用 t 检验完成各组均值间的配对比较，对多重比较误差率不进行调整。
- 邦弗伦尼 (Bonferroni)：用 t 检验完成各组间均值的配对检验，但通过设置每个检验的误差率来控制整个误差率。

- 斯达克 (Sidak): 基于 t 统计量进行多重配对比较, 可以调整显著性水平。
- 雪费 (Scheffe): 使用样本的 F 分布, 对所有可能的均值组合进行同步配对比较。
- R-E-G-W F: 基于 F 检验的多重比较。
- R-E-G-W Q: 基于学生化范围的多重比较。
- S-N-K: 使用学生化范围分布进行组间均值的配对比较。
- 图基 (Tukey): 使用学生化范围统计量进行组间均值的配对比较。
- 图基 s-b (Tukeys-b): 使用学生化范围分布进行组间均值的配对比较。
- 邓肯 (Duncan): 使用与 SNK 检验相似的逐步过程进行多重比较。
- 霍赫伯格 GT2 (Hochberg's GT2): 使用学生化最大系数进行多重比较。
- 加布里埃尔 (Gabriel): 使用学生化最大系数进行配对比较。
- 沃勒-邓肯 (Waller-Duncan): 使用 t 统计量进行多重比较检验。
- 邓尼特 (Dunnett): 指定一个控制组, 其他组都与控制组进行多重配对 t 检验。



图 9-2 “对比 (Contrasts) 设置”对话框



图 9-3 “事后比较 (Post Hoc)” 设置对话框

(2) 不假定等方差 (Equal Variances Not Assumed) 选项栏

用于定义当样本方差不齐次情况下多重比较的检验方法。

- 塔姆黑尼 T2 (Tamhane's T2): 基于 t 检验的保守的配对比较。
- 邓尼特 T3 (Dunnett's T3): 基于学生化最大系数的配对比较。
- 盖姆斯-豪厄尔 (Games-Howell): 是方差不齐次时的一种比较灵活的配对比较。
- 邓尼特 C (Dunnett's C): 该方法是基于学生化范围的配对检验。

(3) 显著性水平 (Significance Level)

定义事后比较的显著性水平。

4. 选项 (Options) 设置

单击“选项 (Options)”按钮, 则弹出如图 9-4 所示对话框, 此对话框用于设置输出选项及缺失值的处理方式。



图 9-4 “选项 (Options) 设置”对话框

统计 (Statistics) 选项栏, 此栏用于选择哪些统计量。

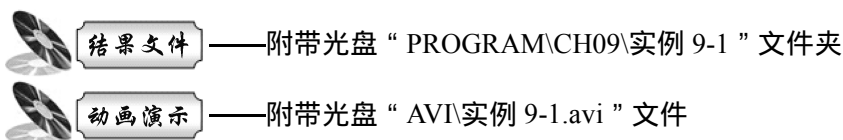
- 描述性 (Descriptive): 描述性统计量。
- 固定和随机效应 (Fixed and Random Effects): 包括固定效应模式和随机效应模式的相关统计量。
- 方差同质性检验 (Homogeneity of Variance Test): 方差齐次性检验结果。
- Brown-Forsythe: 检验各组均值是否相等的检验统计量。
- Welch: 检验各组均值是否相等的检验统计量。

平均值图 (Means plot): 输出均值分布图。

缺失值 (Missing Values) 选项栏: 用于设置缺失值的处理方法。

- 按具体分析排除个案 (Export cases analysis by analysis): 剔除含有缺失值的观测;
- 成列排除个案 (Export cases listwise): 当某条记录有一个因素变量或因变量含有缺失值, 则剔除这条观测记录。

9.2.4 实例分析



本实例所用数据为 SPSS 自带的数据集 salesperformance.sav, 数据集为三组销售绩效数据, 下面就对这三组数据进行方差分析, 数据格式如图 9-5 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|---------|----|----|------|---------------------|---|----|---|----|----|----|
| 1 | group | 数字 | 1 | 0 | Sales training g... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 2 | perform | 数字 | 8 | 2 | Score on trainin... | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 9-5 数据集 salesperformance.sav 的数据格式

1. 参数设置

选择菜单“分析 (Analyze) 比较平均值 (Compare Means) 单因素 ANOVA 检验 (One-Way ANOVA)”, 则弹出如图 9-6 所示对话框, 选择变量“Score on training exam”到“因变量列表 (Dependent List)”选项栏, 选择变量“Sales training group”到“因子 (Factor)”选项栏。

单击“选项 (Options)”按钮则弹出如图 9-4 所示对话框, 选中“描述性”选项栏和“方差同质性”检验选项栏, 然后单击“继续”按钮返回主界面。



图 9-6 “单因素 ANOVA 检验 (One-Way ANOVA) 设置”对话框

2. 结果分析

单击主界面单因素 ANOVA 检验 (One-Way ANOVA) 中的“确定 (OK)”按钮, 则进行分析。结果如图 9-7 所示是方差齐性检验结果。Levene 统计量取值为 4.637, Sig 取值为 0.014, 小于 0.05, 所以认为各组的方差没有齐性。

| 方差齐性检验 | | | |
|------------------------|-------|-------|------|
| Score on training exam | | | |
| 莱文统计 | 自由度 1 | 自由度 2 | 显著性 |
| 4.637 | 2 | 57 | .014 |

图 9-7 方差齐性检验结果

如图 9-8 所示的是方差分析表。分别是平方和、自由度、均方和、 F 值和显著性指标。Sig 取值为 0.000, 小于 0.05, 即假设不成立, 认为各组均值有差异性。

| ANOVA | | | | | |
|------------------------|----------|-----|----------|--------|------|
| Score on training exam | | | | | |
| | 平方和 | 自由度 | 均方 | F | 显著性 |
| 组间 | 2525.691 | 2 | 1262.846 | 12.048 | .000 |
| 组内 | 5974.724 | 57 | 104.820 | | |
| 总计 | 8500.415 | 59 | | | |

图 9-8 方差分析表

9.3 多因素方差分析

在现实中, 常常会遇到两个因素同时影响结果的情况。这就需要检验究竟是一个因素起作用, 还是两个因素都起作用, 或者两个因素的影响都不显著。

双因素方差分析有两种类型：一种是无交互作用的双因素方差分析，它假设因素 A 和因素 B 的效应之间是相互独立的，不存在相互关系；另一种是有交互作用的方差分析，它假设 A 、 B 两个因素不是独立的，而是相互起作用的，两个因素同时起作用的结果不是两个因素分别作用的简单相加，两者的结合会产生一个新的效应。这种效应的最典型的例子是，耕地深度和施肥量都会影响产量，但同时深耕和适当的施肥可能使产量成倍增加，这时，耕地深度和施肥量就存在交互作用。两个因素结合后就会产生出一个新的效应，属于有交互作用的方差分析问题。

9.3.1 只考虑主效应的多因素方差分析

设两个因素分别是 A 和 B 。因素 A 共有 r 个水平，因素 B 共有 s 个水平，无交互作用的双因素方差分析的数据结构参见表 9-3。

表 9-3 无交互作用双因素方差分析的数据结构

| j | | 因 素 B | | | | |
|-------------------|----------|---------------------|---------------------|----------|---------------------|--------------------|
| | | B_1 | B_2 | ... | B_s | 均值 |
| 因 素 A | A_1 | x_{11} | x_{12} | ... | x_{1s} | $\bar{x}_{1\cdot}$ |
| | A_2 | x_{21} | x_{22} | ... | x_{2s} | $\bar{x}_{2\cdot}$ |
| | \vdots | \vdots | \vdots | \vdots | \vdots | \vdots |
| | A_r | x_{r1} | x_{r2} | ... | x_{rs} | $\bar{x}_{r\cdot}$ |
| | 均值 | $\bar{x}_{\cdot 1}$ | $\bar{x}_{\cdot 2}$ | ... | $\bar{x}_{\cdot s}$ | |

方差分析模型为

$$x_{ij} = \mu + a_i + \beta_j + \varepsilon_{ij}$$

式中， x_{ij} 表示第组的第 j 个观察值； μ 表示总体的平均水平； a_i 表示影响因素 A 在 i 水平下对应变量的附加效应； β_j 表示影响因素 B 在 j 水平下对应变量的附加效应，并满足

$$\sum_{j=1}^r \alpha_j = 0 \text{ 和 } \sum_{i=1}^s \beta_i = 0$$

ε_{ij} 为一个服从正态分布 $N(0, \sigma^2)$ 的随机变量，代表随机误差。检验因素 A 是否起作用实际上就是检验各个 a_j 是否均为 0，如都为 0，则因素 A 所对应的各组总体均数都相等，即因素 A 的作用不显著；对因素 B 也是这样。因此原假设有两个。

对因素 A ： $H_{01} : \alpha_i = 0$ ； $H_{11} : \alpha_i$ 不全为 0，等价于 $H_{01} : \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{r\cdot}$ ； $H_{11} : \mu_{1\cdot}, \mu_{2\cdot}, \cdots, \mu_{r\cdot}$ 不全等。

对因素 B ： $H_{02} : \beta_j = 0$ ； $H_{12} : \beta_j$ 不全为 0，等价于 $H_{02} : \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot s}$ ； $H_{12} : \mu_{\cdot 1}, \mu_{\cdot 2}, \cdots, \mu_{\cdot s}$ 不全等。

1. 检验 F 统计量

(1) 水平的均值

$$\bar{x}_{i\cdot} = \frac{\sum_{j=1}^s x_{ij}}{s} ; \bar{x}_{\cdot j} = \frac{\sum_{i=1}^r x_{ij}}{r}$$

(2) 总均值

$$\bar{\bar{x}} = \frac{\sum_{i=1}^r \sum_{j=1}^s x_{ij}}{rs} = \frac{\sum_{i=1}^r \bar{x}_{i\cdot}}{r} = \frac{\sum_{j=1}^s \bar{x}_{\cdot j}}{s}$$

(3) 离差平方和的分解

双因素方差分析同样要对总离差平方和 SST 进行分解, SST 分解为三部分: SSA、SSB 和 SSE, 以分别反映因素 A 的组间差异、因素 B 的组间差异和随机误差的离散状况。它们的计算公式分别为

$$SST = \sum_{i=1}^r \sum_{j=1}^s (x_{ij} - \bar{\bar{x}})^2 ; SSA = \sum_{i=1}^r s(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 ; SSB = \sum_{j=1}^s r(\bar{x}_{\cdot j} - \bar{\bar{x}})^2$$

$$SSE = SST - SSA - SSB$$

(4) 构造检验统计量

由平方和与自由度可以计算出均方, 从而计算出 F 检验值, 参见表 9-4。

表 9-4 无交互作用的双方差分析表

| 方差来源 | 离差平方和 | df | 均方 (MS) | F |
|--------|-------|--------------|---------------------------------|-----------|
| 因素 A | SSA | $r-1$ | $MSA = SSA / (r-1)$ (9.23) | MSA/MSE |
| 因素 B | SSB | $s-1$ | $MSB = SSE / (n-r)$ (9.24) | MSB/MSE |
| 误差 | SSE | $(r-1)(s-1)$ | $MSE = SSE / (r-1)(s-1)$ (9.25) | |
| 总方差 | SST | $n-1$ | | |

为检验因素 A 的影响是否显著, 采用下面的统计量:

$$F_A = \frac{MSA}{MSE} \sim F_{\alpha}(r-1, n-r-s+1)$$

为检验因素 B 的影响是否显著, 采用下面的统计量:

$$F_B = \frac{MSB}{MSE} \sim F_{\alpha}(s-1, n-r-s+1)$$

2. 判断与结论

根据给定的显著性水平 α 在 F 分布表中查找相应的临界值 F_{α} , 将统计量 F 与 F_{α} 进行比较, 做出拒绝或不能拒绝原假设 H_0 的决策。

若 $F_A \geq F_{\alpha}$, 则拒绝原假设 H_{01} , 表明均值之间有显著差异, 即因素 A 对观察值有显著影响;

若 $F_A < F_{\alpha}$, 则不能拒绝原假设 H_{01} , 表明均值之间的差异不显著, 即因素 A 对观察值

没有显著影响；

若 $F_B > F_\alpha$ ，则拒绝原假设 H_{02} ，表明均值之间有显著差异，即因素 B 对观察值有显著影响；

若 $F_B < F_\alpha$ ，则不能拒绝原假设 H_{02} ，表明均值之间的差异不显著，即因素 B 对观察值没有显著影响。

9.3.2 存在交互效应的多因素方差分析

设两个因素分别是 A 和 B ，因素 A 共有 r 个水平，因素 B 共有 s 个水平，若对两个因素的交互作用进行分析，每组试验条件的试验至少要进行两次，若对每个水平组合水平下 (A_j, B_i) 重复 t 次试验，每次试验的结果用 x_{ijk} 表示，那么有交互作用的双因素方差分析的数据结构参见表 9-5。

表 9-5 有交互作用双因素方差分析的数据结构

| $j \backslash i$ | | 因 素 B | | | |
|-------------------|----------|------------------------------------|----------|------------------------------------|----------------------|
| | | B_1 | ... | B_s | 均值 |
| 因 素 A | A_1 | $x_{111}, x_{112}, \dots, x_{11t}$ | ... | $x_{1s1}, x_{1s2}, \dots, x_{1st}$ | $\bar{x}_{1\bullet}$ |
| | A_2 | $x_{211}, x_{212}, \dots, x_{21t}$ | ... | $x_{2s1}, x_{2s2}, \dots, x_{2st}$ | $\bar{x}_{2\bullet}$ |
| | \vdots | \vdots | \vdots | \vdots | \vdots |
| | A_r | $x_{r11}, x_{r12}, \dots, x_{r1t}$ | ... | $x_{rs1}, x_{rs2}, \dots, x_{rst}$ | $\bar{x}_{r\bullet}$ |
| | 均值 | $\bar{x}_{\bullet 1}$ | | $\bar{x}_{\bullet s}$ | |

方差分析模型如下：

$$x_{ijk} = \mu + a_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

式中， x_{ijk} 表示第 ij 组中的第 k 个观察值； μ 表示总体的平均水平； a_i 表示影响因素 A 在 i 水平下对应变量的附加效应； β_j 表示影响因素 B 在 j 水平下对应变量的附加效应， $(\alpha\beta)_{ji}$ 为两者的交互效应，并满足 $\sum_{j=1}^r \alpha_j = 0$ 、 $\sum_{i=1}^s \beta_i = 0$ 和 $\sum_{j=1}^r (\alpha\beta)_{ij} = 0$ 、 $\sum_{i=1}^s (\alpha\beta)_{ij} = 0$ ； ε_{ijk} 为一个服从正态分布 $N(0, \sigma^2)$ 的随机变量。与前面的分析思路相同，检验因素 A 、因素 B ，以及两者的交互效应是否起作用实际上就是检验各个 a_j 、 β_i ，以及 $(\alpha\beta)_{ij}$ 是否均为 0。故原假设有三个。

对因素 A ： $H_{01} : \alpha_i = 0$ ； $H_{11} : \alpha_i$ 不全为 0。

对因素 B ： $H_{02} : \beta_j = 0$ ； $H_{12} : \beta_j$ 不全为 0。

对因素 A 和 B 的交互效应： $H_{03} : (\alpha\beta)_{ij} = 0$ ； $H_{13} : (\alpha\beta)_{ij}$ 不全为零。

1. 构造检验 F 统计量

(1) 水平的均值

$$\bar{x}_{ij} = \frac{\sum_{k=1}^t x_{ijk}}{t} ; \bar{x}_{\bullet i} = \frac{\sum_{j=1}^s \sum_{k=1}^t x_{ijk}}{st} ; \bar{x}_{j\bullet} = \frac{\sum_{i=1}^r \sum_{k=1}^t x_{ijk}}{rt}$$

(2) 总均值

$$\bar{\bar{x}} = \frac{\sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t x_{ijk}}{rst} = \frac{\sum_{r=1}^r \bar{x}_{\bullet r}}{r} = \frac{\sum_{j=1}^s \bar{x}_{j\bullet}}{s}$$

(3) 离差平方和的分解

与无交互作用的双因素方差分析不同, 总离差平方和 SST 将被分解为 4 个部分: SSA、SSB、SSAB 和 SSE, 以分别反映因素 A 的组间差异、因素 B 的组间差异、因素 AB 的交互效应和随机误差的离散状况。

它们的计算公式分别为

$$\begin{aligned} \text{SST} &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t (x_{ijk} - \bar{\bar{x}})^2 ; \text{SSA} = \sum_{i=1}^r st(\bar{x}_{\bullet i} - \bar{\bar{x}})^2 ; \text{SSB} = \sum_{j=1}^s rt(\bar{x}_{j\bullet} - \bar{\bar{x}})^2 \\ \text{SSAB} &= \sum_{i=1}^r \sum_{j=1}^s t(\bar{x}_{ij} - \bar{x}_{\bullet i} - \bar{x}_{j\bullet} + \bar{\bar{x}})^2 ; \text{SSE} = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t (x_{ijk} - \bar{x}_{ij})^2 \end{aligned}$$

(4) 构造检验统计量

由平方和与自由度可以计算出均方, 从而计算出 F 检验值, 参见表 9-6。

表 9-6 有交互作用的双方差分析表

| 方 差 来 源 | 离差平方和 | df | 均方 (MS) | F |
|----------|-------|------------|----------------------|----------|
| 因素 A | SSA | r-1 | MSA = SSA / (r-1) | MSA/MSE |
| 因素 B | SSB | s-1 | MSB = SSE / (n-r) | MSB/MSE |
| 因素 A × B | SSAB | (r-1)(s-1) | MSAB=SSAB/(r-1)(s-1) | MSAB/MSE |
| 误差 | SSE | rs(t-1) | MSE= SSE / rs(t-1) | |
| 总方差 | SST | n-1 | | |

为检验因素 A 的影响是否显著, 采用下面的统计量, 即

$$F_A = \frac{\text{MSA}}{\text{MSE}} \sim F_{\alpha}(r-1, n-rs)$$

为检验因素 B 的影响是否显著, 采用下面的统计量, 即

$$F_B = \frac{\text{MSB}}{\text{MSE}} \sim F_{\alpha}(s-1, n-rs)$$

为检验因素 A、B 交互效应的影响是否显著, 采用下面的统计量, 即

$$F_{AB} = \frac{\text{MSAB}}{\text{MSE}} \sim F_{\alpha}(n-r-s+1, n-rs)$$

2. 判断与结论

根据给定的显著性水平 α 在 F 分布表中查找相应的临界值 F_{α} , 将统计量 F 与 F_{α} 进行比

较，做出拒绝或不能拒绝原假设 H_0 的决策：

若 $F_A = F_\alpha(r-1, n-rs)$ ，则拒绝原假设 H_{01} ，表明因素 A 对观察值有显著影响；

若 $F_B = F_\alpha(s-1, n-rs)$ ，则拒绝原假设 H_{02} ，表明因素 B 对观察值有显著影响；

若 $F_{AB} = F_\alpha(n-r-s+1, n-rs)$ ，则拒绝原假设 H_{03} ，表明因素 A 、 B 的交互效应对观察值有显著影响。

9.3.3 单变量过程参数设置

选择菜单“分析 (Analyze) — 一般线性模型 (General Linear Model) — 单变量 (Univariate)”，则弹出如图 9-9 所示对话框。此对话框用于设置多重方差分析的各种参数，各部分组成如下所述。

1. 变量选择设置

图 9-9 左边是待分析的变量框。



图 9-9 “单变量 (Univariate) 过程参数设置”对话框

- 因变量 (Dependent Variable)：此栏用于选择指标变量。
- 固定因子 (Fixed Factor (s))：选择固定因素变量。
- 随机因子 (Random Factor (s))：选择随机因素变量。
- 协变量 (Covariate (s))：选择协变量，用于协方差分析。
- WLS 权重 (WLS Weight)：选择加权最小二乘法的权重系数。

2. 模型 (Model) 设置

单击图 9-9 中的“模型 (Model)”按钮，则弹出如图 9-10 所示对话框，此对话框用于设置方差分析的模型。

全因子 (Full Factor) 选项框：系统默认项，用于建立全模型，分析所有因素的主效应及其交互效应。



图 9-10 “模型 (Model) 设置”对话框

定制 (Custom) 选项框：用户自定义反差分析模型，选择后则激活其下的选项框和下拉菜单。

- 因子与协变量 (Factors & Covariates)：列出在 Univariate 过程中选择的所有的固定因素变量 (F)、随机因素变量 (R) 和协变量 (C)。
- 模型 (Model)：选择方差分析的主效应。若同时将因子与协变量 (Factors & Covariates) 选项框中两个变量选入，则将其交互效应强行纳入模型。

构建项 (Build Term(s)) 下拉菜单。

- 交互 (Interaction)：定义进行选择变量的交互效应的方差分析。
- 主效应 (Main Effects)：定义进行选择变量的主效应的方差分析。
- All 2-way—All 5-way：定义进行所有变量的 i 阶交互效应的方差分析。

平方和 (Sum of Squares)：定义平方和的分解方法。

在模型中包含截距 (Include Intercept in Model)：用于选择模型中是否包含截距平方和。

3. 对比 (Contrasts) 设置

单击图 9-9 中的“对比 (Contrasts)”按钮，则弹出如图 9-11 所示对话框，此对话框用于设置比较各因素水平之间的差异。



图 9-11 “对比 (Contrasts) 设置”对话框

因子 (Factor (s)) 选项框：列出所有因素变量。

更改对比 (Contrast) 下拉菜单：用于设置比较因素水平间差异的方法，各种方法如下：

- 无 (None)
- 偏差比较法 (Deviation)
- 简单比较法 (Simple)
- 差值比较法 (Difference)
- 赫尔默特比较法 (Helmert)
- 重复比较法 (Repeated)
- 多项式比较法 (Polynomial)

变化量 (Change)：单击此按钮，改变选中因素的比较方法。

4. 图 (Plots) 设置

单击图 9-9 中的“图”按钮，则弹出如图 9-12 所示对话框，此对话框用于设置图形的输出参数。



图 9-12 “图设置”对话框

因子 (Factor (s))：列举可用于作图的变量。

水平轴 (Horizontal Axis)：选择作为横坐标的因变量。

单轴的线条 (Separate Lines)：选择曲线分组变量，按照该变量的不同取值在同一张图上绘制多条曲线。

单独的图：选择图形分组变量，按照该变量的不同取值绘制多张图形。

图：用于添加 (Add)、更改 (Change)、除去 (Remove) 已经定义的图形。

5. 事后比较 (Post Hoc) 设置

单击图 9-9 中的“事后比较 (Post Hoc)”按钮，则弹出如图 9-13 所示对话框，此对话框用于设置各因素方差分析多重比较的检验方法。此对话框与单因素 ANOVA 检验过程中的事后比较 (Post Hoc) 选项框完全类似，在此不再赘述。



图 9-13 “事后比较 (Post Hoc) 选项设置”对话框

6. 保存 (Save) 设置

单击图 9-9 中的“保存 (Save)”按钮, 则弹出如图 9-14 所示对话框, 此对话框用于设置保存分析中的结果。

预测值 (Predicted Values) 选项: 用于设置预测值。

- 未标准化 (Unstandardized): 没有标准化的预测值。
- 加权 (Weighted): 加权预测值。
- 标准误差 (Standard Error): 没有标准化预测值的标准误差。

诊断 (Diagnostics) 选项: 诊断方法。

- CoOK 距离 (CoOK's Distance): 库克距离。
- 杠杆值 (Leverage Values): 非中心化杠杆值。

残差 (Residuals) 选项: 残差选项。

- 未标准化 (Unstandardized): 没有标准化的残差。
- 加权 (Weighted): 加权残差。
- 标准化 (Standardized): 标准化残差。
- 学生化 (Studentized): 学生化残差。
- 删除后 (Deleted): 删除残差。

系数统计 (Coefficient Statistics) 选项: 此选项框用于保存参数拟合的协方差矩阵。

- 创建新数据集 (Creates New Dataset): 保存在一个新的文件中。
- 写入新数据文件 (Write a New Data File): 直接保存到一个其他的文件中。

7. 选项 (Options) 设置

单击图 9-9 中的“选项 (Options)”按钮, 则弹出如图 9-15 所示对话框, 此对话框用于进行选项设置, 包括输出、估计边际均值选项。

因子与因子交互 (Factor(s) and Factor Interactions): 列出可选的因素变量及其交互作用, 其中的 OVERALL 代表对所有的因素及其交互作用都计算其对应的样本均值。

显示均值 (Display Means for): 将因子与因子交互选项栏中要计算的变量选入此栏中。

比较主效应 (Compare Main effects): 当显示均值选项框中有元素时, 该选项被激活, 用来定义是否对选中变量进行均值的多重比较。

置信区间调整 (Confidence Interval Adjustment): 选择多重比较的迭代。

显示 (Display) 选项: 用于定义输出统计量。

- 描述统计 (Descriptive Statistics): 描述性统计量。
- 效率量估算 (Estimate of Effect Size): 计算因素偏差。
- 实测幂 (Observed Power): 功效检验。
- 参数估算值 (Parameter Estimate): 将各因素水平转化为哑变量之后估计其多元线性模型的系数。
- 对比系数矩阵 (Contrast Coefficient Matrix)。
- 齐性检验 (Homogeneity Tests): 水平间的方差齐次性检验。
- 分布-水平图 (Spread vs. Level Plot): 绘制单元格的均值对于标准差、方差的散点图。
- 残差图 (Residual Plot): 绘制预测值、观察值, 以及残差间的散点图。
- 失拟 (Lack of Fit): 检查当前模型是否能够合理描述自变量和因变量之间的关系。
- 一般可估函数 (General Estimate Function): 显示估计函数的通用表格。

显著性水平 (Significance Level): 定义显著性水平。




图 9-14 “保存 (Save) 设置”对话框



图 9-15 “选项 (Options) 设置”对话框

9.3.4 实例分析

 **结果文件** —— 附带光盘 “PROGRAM\CH09\实例 9-2” 文件夹

 **动画演示** —— 附带光盘 “AVI\实例 9-2.avi” 文件

本实例所用的是 SPSS 自带的数据集 grocery_1month.sav, 此数据集有 14 个变量, 如变量 storeid (商店号码)、hlthfood (健康食品商店)、size (商店大小)、custid (客户代码)。

gender (性别) shopfor (购物目的) style (购物方式) 等, 数据集有 351 个观测样本数。本调查数据的数据格式如图 9-16 所示。

1. 参数设置

选择菜单“分析 (Analyze) — 一般线性模型 (General Linear Model) — 单变量 (Univariate)”, 则弹出如图 9-17 所示对话框。此对话框用于设置多重方差分析的各种参数。选中变量 Amount spent 到“因变量 (Dependent Variable)”选项栏, 选中变量 Gender 和 Shopping style 到“固定因子 (fixed Factors)”选项栏。

然后单击“图”按钮, 弹出如图 9-18 所示对话框, 选中变量 style 到“水平轴 (Horizontal Axis)”选项栏中, 选中变量 gender 到“单图 (Separate Lines)”选项栏中, 然后单击“添加 (Add)”按钮, 设置后单击“继续”按钮返回主界面。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|----|-----------|----|----|------|--------------------|------------------|----|---|----|----|----|
| 1 | storeid | 数字 | 4 | 0 | Store ID | 无 | 无 | 8 | 右 | 名义 | 输入 |
| 2 | hlthfood | 数字 | 4 | 0 | Health food store | {0, No}... | 无 | 8 | 右 | 名义 | 输入 |
| 3 | size | 数字 | 4 | 0 | Size of store | {1, Small}... | 无 | 8 | 右 | 有序 | 输入 |
| 4 | org | 数字 | 4 | 0 | Store organization | {1, Emphasi}... | 无 | 8 | 右 | 名义 | 输入 |
| 5 | custid | 数字 | 4 | 0 | Customer ID | 无 | 无 | 8 | 右 | 名义 | 输入 |
| 6 | gender | 数字 | 4 | 0 | Gender | {0, Male}... | 无 | 8 | 右 | 名义 | 输入 |
| 7 | shopfor | 数字 | 4 | 0 | Who shopping for | {1, Self}... | 无 | 8 | 右 | 名义 | 输入 |
| 8 | veg | 数字 | 4 | 0 | Vegetarian | {0, No}... | 无 | 8 | 右 | 名义 | 输入 |
| 9 | style | 数字 | 4 | 0 | Shopping style | {1, Biweekly}... | 无 | 8 | 右 | 名义 | 输入 |
| 10 | usecoup | 数字 | 4 | 0 | Use coupons | {1, No}... | 无 | 8 | 右 | 名义 | 输入 |
| 11 | amtspend | 数字 | 8 | 2 | Amount spent | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 12 | pre_1 | 数字 | 8 | 2 | Predicted Value | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 13 | qcl_1 | 数字 | 8 | 0 | Cluster Number | 无 | 无 | 8 | 右 | 名义 | 输入 |
| 14 | filter_\$ | 数字 | 1 | 0 | qcl_1 = 3 (FILTER) | {0, Not Sele}... | 无 | 8 | 右 | 标度 | 输入 |

图 9-16 数据集 grocery_1month.sav 的数据格式



图 9-17 “单变量 (Univariate) 设置”对话框



图 9-18 “图设置”对话框

然后单击“事后比较 (Post Hoc)”按钮, 打开如图 9-19 所示对话框。选中变量 style 到“事后比较检验 (Post Hoc tests for)”选项栏中, 选中 Tukey 选项栏。设置后单击“继续”

(Continue)”按钮返回主界面。

接着单击“选项 (Options)”按钮,弹出如图 9-20 所示对话框。选中变量 gender*style 到“显示平均值 (Display Means for)”选项栏中,选中“描述统计 (Descriptive Statistics)”、“齐性检验 (Homogeneity Tests)”、“效率量估算 (Estimates of effect size)”,以及“分布-水平图 (Spread vs. level plot)”选项栏,设置后单击“继续”按钮返回主界面。

2. 结果分析

设置参数完成后单击主界面中的“确定”按钮进行分析,结果如下。如图 9-21 所示是描述性统计信息,给出了分组的均值、标准差,以及样本数。从图中可以看出变量 Shopping style 平均情况下“biweekly”的顾客消费 378.5210,“weekly”的顾客消费 404.5552,“often”的顾客消费 406.76,也可以得到性别不同情况下的平均消费情况。



图 9-19 “事后比较 (Post Hoc) 设置”对话框



图 9-20 “选项 (Options) 设置”对话框

如图 9-22 所示的是 Levene 检验结果,由显著性检验的 Sig 值 0.330 大于 0.10,所以在 0.10 的显著性水平上,认为各组方差是无显著性差异的。

| 描述统计 | | | | |
|-------------------|-----------------------|----------|-----------|-----|
| 因变量: Amount spent | | | | |
| Gender | Shopping style | 平均值 | 标准偏差 | 个案数 |
| Male | Biweekly; in bulk | 413.0657 | 90.86574 | 35 |
| | Weekly; similar items | 440.9647 | 98.23860 | 120 |
| | Often; what's on sale | 407.7747 | 69.33334 | 30 |
| | 总计 | 430.3043 | 93.47877 | 185 |
| Female | Biweekly; in bulk | 343.9763 | 100.47207 | 35 |
| | Weekly; similar items | 361.7205 | 90.46076 | 102 |
| | Often; what's on sale | 405.7269 | 80.57058 | 29 |
| | 总计 | 365.6671 | 92.64058 | 166 |
| 总计 | Biweekly; in bulk | 378.5210 | 101.25839 | 70 |
| | Weekly; similar items | 404.5552 | 102.48440 | 222 |
| | Often; what's on sale | 406.7681 | 74.42114 | 59 |
| | 总计 | 399.7352 | 98.40821 | 351 |

图 9-21 描述性统计信息

| 误差方差的莱文等同性检验 ^a | | | |
|---|-------|-------|------|
| 因变量: Amount spent | | | |
| F | 自由度 1 | 自由度 2 | 显著性 |
| 1.157 | 5 | 345 | .330 |
| 检验“各个组中的因变量误差方差相等”这一原假设。 | | | |
| a. 设计: 截距 + gender + style + gender * style | | | |

图 9-22 Levene 检验结果

如图 9-23 所示的是效应检验结果,除了 gender*style 的交互效应外,其他的变量对消

费额的影响都是显著的。

| 主体间效应检验 | | | | | | |
|-------------------|-------------------------|-----|-------------|----------|------|----------|
| 因变量: Amount spent | | | | | | |
| 源 | III 类平方和 | 自由度 | 均方 | F | 显著性 | 偏 Eta 平方 |
| 修正模型 | 469402.996 ^a | 5 | 93880.599 | 11.092 | .000 | .138 |
| 截距 | 39359636.39 | 1 | 39359636.39 | 4650.274 | .000 | .931 |
| gender | 158037.442 | 1 | 158037.442 | 18.672 | .000 | .051 |
| style | 33506.210 | 2 | 16753.105 | 1.979 | .140 | .011 |
| gender * style | 69858.325 | 2 | 34929.163 | 4.127 | .017 | .023 |
| 误差 | 2920058.824 | 345 | 8463.939 | | | |
| 总计 | 59475118.44 | 351 | | | | |
| 修正后总计 | 3389461.820 | 350 | | | | |

a. R 方 = .138 (调整后 R 方 = .126)

图 9-23 效应检验结果

图 9-24 是参数估计结果,给出了均值、标准误,统计量检验 P 值,以及置信区间为 95% 的上限和下限。

下面是消费额的分布和水平图,如图 9-25 所示,是关于标准差的分布和水平图。

最后输出的是边际均值图,如图 9-26 所示,图中显示的男性和女性的折线没有交叉,说明他们之间的消费额差异比较显著。

| 多重比较 | | | | | | |
|-----------------------|-----------------------|-------------|----------|------|----------|---------|
| 因变量: Amount spent | | | | | | |
| 图基 HSD | | | | | | |
| (I) Shopping style | (J) Shopping style | 平均值差值 (I-J) | 标准误差 | 显著性 | 95% 置信区间 | |
| Biweekly; in bulk | Weekly; similar items | -26.0342 | 12.61108 | .099 | -55.7191 | 3.6507 |
| | Often; what's on sale | -28.2471 | 16.25946 | .193 | -66.5198 | 10.0256 |
| Weekly; similar items | Biweekly; in bulk | 26.0342 | 12.61108 | .099 | -3.6507 | 55.7191 |
| | Often; what's on sale | -2.2130 | 13.47525 | .985 | -33.9320 | 29.5061 |
| Often; what's on sale | Biweekly; in bulk | 28.2471 | 16.25946 | .193 | -10.0256 | 66.5198 |
| | Weekly; similar items | 2.2130 | 13.47525 | .985 | -29.5061 | 33.9320 |

基于实测平均值。
误差项是均方(误差) = 8463.939。

图 9-24 参数估计结果

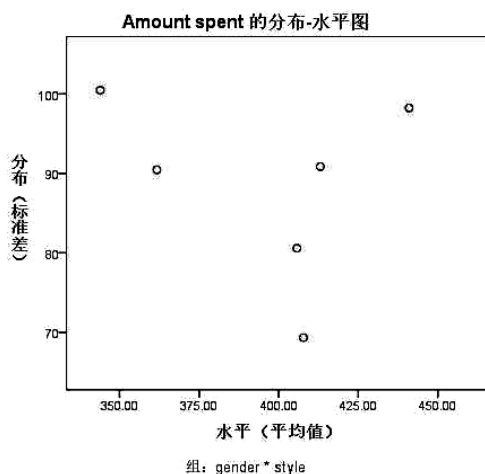


图 9-25 消费额的分布和水平图

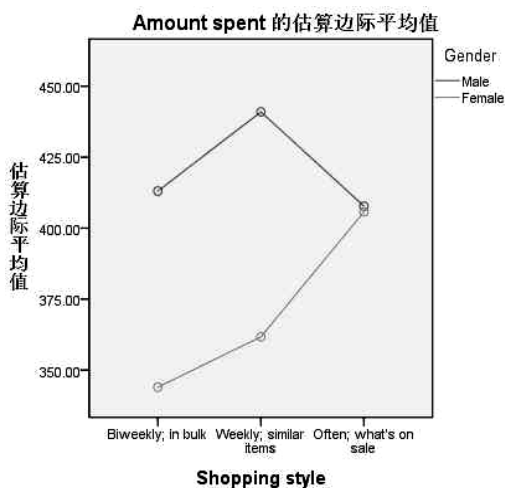


图 9-26 边际均值图

9.4 协方差分析

协方差分析是建立在方差分析和回归分析基础之上的一种统计分析方法。方差分析是从质量因子的角度探讨因素不同水平对实验指标影响的差异。一般说来，质量因子是可以人为控制的。

9.4.1 协方差分析概述

两个相关变量线性相关性质与程度的相关系数的计算公式如下。

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

若将公式右端的分子分母同除以自由度 $(n-1)$ ，得

$$r = \frac{\sum (x - \bar{x})(y - \bar{y}) / (n-1)}{\sqrt{\left[\sum (x - \bar{x})^2 / (n-1) \right] \left[\sum (y - \bar{y})^2 / (n-1) \right]}}$$

式中， $\frac{\sum (x - \bar{x})^2}{n-1}$ 是 x 的均方 MS_x ，它是 x 的方差 σ_x^2 的无偏估计量；

$\frac{\sum (y - \bar{y})^2}{n-1}$ 是 y 的均方 MS_y ，它是 y 的方差 σ_y^2 的无偏估计量；

$\frac{\sum (x - \bar{x})(y - \bar{y})}{n-1}$ 称为 x 与 y 的平均离均差的乘积和，简称均积，记为 MP_{xy} ，即

$$MP_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{n-1}$$

与均积相应的总体参数称为协方差 (Covariance), 记为 $\text{Cov}(x, y)$ 或 σ_{xy} 。统计学证明, 均积 MP_{xy} 是总体协方差 $\text{Cov}(x, y)$ 的无偏估计量, 即 $\text{EMP}_{xy} = \text{Cov}(x, y)$ 。

于是, 样本相关系数 r 可用均方 MS_x 、 MS_y 、均积 MP_{xy} 表示为

$$r = \frac{\text{MP}_{xy}}{\sqrt{\text{MS}_x \text{MS}_y}}$$

相应的总体相关系数 ρ 可用 x 与 y 的总体标准差 σ_x 、 σ_y , 总体协方差 $\text{Cov}(x, y)$ 或 σ_{xy} 表示为

$$\rho = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

均积与均方具有相似的形式, 也有相似的性质。在方差分析中, 一个变量的总平方和与自由度可按变异来源进行剖分, 从而求得相应的均方。统计学已证明: 两个变量的总乘积和与自由度也可按变异来源进行剖分而获得相应的均积。这种把两个变量的总乘积和与自由度按变异来源进行剖分并获得相应均积的方法亦称为协方差分析。

在随机模型的方差分析中, 根据均方 MS 和期望均方 EMS 的关系, 可以得到不同变异来源的方差组分的估计值。同样, 在随机模型的协方差分析中, 根据均积 MP 和期望均积 EMP 的关系, 可得到不同变异来源的协方差组分的估计值。有了这些估计值, 就可进行相应的总体相关分析。这些分析在遗传、育种和生态、环保的研究上是很有用处的。

9.4.2 实例分析



结果文件 —— 附带光盘 “PROGRAM\CH09\实例 9-3” 文件夹



动画演示 —— 附带光盘 “AVI\实例 9-3.avi” 文件

本实例所用的是 SPSS 自带的数据集 workprog.sav, 此数据集有 8 个变量, 即 age、marital、incbef、ed、gender、reside、prog, 数据集有 1000 个观测样本数。本调查数据是参加政府项目前后人们工作、薪金、婚姻、教育水平等信息, 数据如图 9-27 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|---------|----|----|------|--------------------|------------------|----|----|----|----|----|
| 1 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 2 | marital | 数字 | 4 | 0 | Marital status | {0, Unmarrie... | 无 | 7 | 右 | 名义 | 输入 |
| 3 | incbef | 数字 | 8 | 2 | Income before t... | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 4 | incaft | 数字 | 8 | 2 | Income after th... | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 5 | ed | 数字 | 4 | 0 | Level of education | {1, Did not c... | 无 | 6 | 右 | 有序 | 输入 |

图 9-27 数据集 workprog.sav 的格式

1. 参数设置

选择菜单“分析 (Analyze) — 一般线性模型 (General Linear Model) — 单变量 (Univariate)”，则弹出如图 9-28 所示对话框，选择变量 Income after the program 到“因变量 (Dependent Variable)”选项栏，选择变量 Program status 到“固定因子 (Fixed Factor)”选项栏中，选择变量 Income before the program 到“协变量 (Covariate)”选项栏中。

然后单击“模型 (Model)”按钮，弹出如图 9-29 所示对话框，选择“定制 (Custom)”选项栏，首先选中变量 prog 和 incbef，然后选择“构建项 (Build Term(s))”下拉菜单中的“主效应 (Main Effects)”选项，然后选入“模型 (Model)”选项栏中。然后继续选中变量 prog 和 incbef，选择“构建项 (Build Term(s))”下拉菜单中的“交互 (Interaction)”选项，然后选入“模型 (Model)”选项栏中。然后单击“继续 (Continue)”按钮返回主界面。单击主界面的“选项 (Options)”按钮，然后弹出如图 9-30 所示对话框。选中“效率量估算 (Estimates of effect size)”选项栏，再单击“继续 (Continue)”按钮返回主界面。



图 9-28 “单变量 (Univariate) 设置”对话框



图 9-29 “模型 (Model) 设置”对话框



图 9-30 “选项 (Options) 设置”对话框

2. 结果分析

设置好上述参数以后单击主界面中的“确定(OK)”按钮进行分析。首先是基本统计量,如图9-31所示。然后是如图9-32所示的协方差分析结果。从表中可以看出 prog、incbef 所对应的 Sig 取值分别为 0.001、0.000,均小于 0.05,说明变量 prog、incbef 对参加此项目前后变换有显著的影响,但是 prog*incbef 的 Sig 取值为 0.502 大于 0.05,说明此两个变量的交互作用对参加此项目前后变换没有显著的影响。

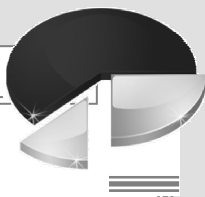
| 主体间因子 | | |
|----------------|---|-----|
| | | 个案数 |
| Program status | 0 | 517 |
| | 1 | 483 |

图 9-31 基本统计量

| 主体间效应检验 | | | | | | |
|-------------------------------|------------------------|------|----------|---------|------|----------|
| 因变量: Income after the program | | | | | | |
| 源 | III 类平方和 | 自由度 | 均方 | F | 显著性 | 偏 Eta 平方 |
| 修正模型 | 12295.033 ^a | 3 | 4098.344 | 429.755 | .000 | .564 |
| 截距 | 131.271 | 1 | 131.271 | 13.765 | .000 | .014 |
| prog | 106.795 | 1 | 106.795 | 11.199 | .001 | .011 |
| incbef | 7152.586 | 1 | 7152.586 | 750.025 | .000 | .430 |
| prog * incbef | 4.292 | 1 | 4.292 | .450 | .502 | .000 |
| 误差 | 9498.318 | 996 | 9.536 | | | |
| 总计 | 297121.000 | 1000 | | | | |
| 修正后总计 | 21793.351 | 999 | | | | |

a. R 方 = .564 (调整后 R 方 = .563)

图 9-32 协方差分析结果



第 10 章 回归分析

回归分析 (Regression Analysis) 是确定两种或两种以上变数间相互依赖的定量关系的一种统计分析方法。本章介绍了线性回归、非线性回归、Logistic 回归, 以及它们对应的 SPSS 过程, 然后结合大量的案例进行研究分析。



本讲内容

- 线性回归
- 非线性回归
- Logistic 回归

10.1 线性回归

在客观世界中, 普遍存在着变量之间的关系数学的一个重要作用就是从数量上来揭示、表达和分析这些关系。而变量之间关系, 一般可分为确定的和非确定的两类, 确定性关系可用函数关系表示, 而非确定性关系则不然。

例如, 人的身高和体重的关系、人的血压和年龄的关系、某产品的广告投入与销售额间的关系等, 它们之间是有关联的, 但是它们之间的关系又不能用普通函数来表示。称这类非确定性关系为相关关系。具有相关关系的变量虽然不具有确定的函数关系, 但是可以借助函数关系来表示它们之间的统计规律, 这种近似地表示它们之间的相关关系的函数称为回归函数。回归分析是研究两个或两个以上变量相关关系的一种重要的统计方法。

在实际中最简单的情形是由两个变量组成的关系。考虑用下列模型表示 $Y = f(x)$, 但是, 由于两个变量之间不存在确定的函数关系, 因此必须把随机波动考虑进去, 故引入模型如下。

$$Y = f(x) + \varepsilon$$

式中, Y 是随机变量, x 是普通变量, ε 是随机变量 (称为随机误差)。

回归分析就是根据已得的试验结果, 以及以往的经验来建立统计模型, 并研究变量间的相关关系, 建立起变量之间关系的近似表达式, 即经验公式, 并由此对相应的变量进行预测和控制等。

10.1.1 线性回归模型

一般地,当随机变量 Y 与普通变量 x 之间有线性关系时,可设

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad (10-1)$$

式中: $\varepsilon \sim N(0, \sigma^2)$; β_0, β_1 为待定系数。

设 $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ 是取自总体 (x, Y) 的一组样本,而 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 是该样本的观察值,在样本和它的观察值中的 x_1, x_2, \dots, x_n 是取定的不完全相同的数值,而样本中的 Y_1, Y_2, \dots, Y_n 在试验前为随机变量,在试验或观测后是具体的数值,一次抽样的结果可以取得 n 对数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 则有

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (10-2)$$

其中 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 相互独立。在线性模型中,由假设知

$$Y \sim N(\beta_0 + \beta_1 x, \sigma^2), \quad E(Y) = \beta_0 + \beta_1 x \quad (10-3)$$

回归分析就是根据样本观察值寻求 β_0, β_1 的估计 $\hat{\beta}_0, \hat{\beta}_1$ 。

对于给定 x 值,取

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (10-4)$$

作为 $E(Y) = \beta_0 + \beta_1 x$ 的估计,式(10-4)称为 Y 关于 x 的线性回归方程或经验公式,其图像称为回归直线, $\hat{\beta}_1$ 称为回归系数。

10.1.2 最小二乘估计

对样本的一组观察值 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 对每个 x_i , 由线性回归方程(10-4)可以确定一回归值,即

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

这个回归值 \hat{y}_i 与实际观察值 y_i 之差,即

$$y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

刻画了 y_i 与回归直线 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 的偏离度,一个自然的想法就是,对所有 x_i 若 y_i 与 \hat{y}_i 的偏离越小,则认为直线与所有试验点拟和得越好。令

$$Q(\beta, \beta) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (10-5)$$

式(10-5)表示所有观察值 y_i 与回归直线 \hat{y}_i 的偏离平方和,刻画了所有观察值与回归直线的偏离度。最小二乘法就是寻求 β_0 与 β_1 的估计 $\hat{\beta}_0, \hat{\beta}_1$, 使 $Q(\hat{\beta}_0, \hat{\beta}_1) = \min Q(\beta_0, \beta_1)$ 。利用微分的方法,求 Q 关于 β_0, β_1 的偏导数,并令其为 0,得

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases}$$

整理得

$$\begin{cases} n\beta_0 + \left(\sum_{i=1}^n x_i\right)\beta_1 = \sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n x_i\right)\beta_0 + \left(\sum_{i=1}^n x_i^2\right)\beta_1 = \sum_{i=1}^n x_i y_i \end{cases}$$

称此为正规方程组，解正规方程组得

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \\ \hat{\beta}_1 = \left(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}\right) / \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) \end{cases}$$

式中： $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ； $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ ，若记

$$L_{xy} \stackrel{\text{def}}{=} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}, \quad L_{xx} \stackrel{\text{def}}{=} \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

则

$$\begin{cases} \hat{\beta}_0 = \hat{y} - \bar{x}\hat{\beta}_1 \\ \hat{\beta}_1 = L_{xy}/L_{xx} \end{cases}$$

上式称为 β_0, β_1 的最小二乘估计。而

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

为 Y 关于 x 的一元经验回归方程。

定理 1：若 $\hat{\beta}_0, \hat{\beta}_1$ 为 β_0, β_1 的最小二乘估计，则 $\hat{\beta}_0, \hat{\beta}_1$ 分别是 β_0, β_1 的无偏估计，且

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right)\right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right)$$

10.1.3 回归方程的显著性检验

前面关于线性回归方程 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 的讨论是在线性假设 $Y = \beta_0 + \beta_1 x + \varepsilon$ ， $\varepsilon \sim N(0, \sigma^2)$ 下进行的。这个线性回归方程是否有实用价值，首先要根据有关专业知识和实践来判断，其次还要根据实际观察得到的数据运用假设检验的方法来判断。

由线性回归模型 $Y = \beta_0 + \beta_1 x + \varepsilon$ ， $\varepsilon \sim N(0, \sigma^2)$ 可知，当 $\beta_1 = 0$ 时，就认为 Y 与 x 之间不存在线性回归关系，故需检验如下假设。

$$H_0: \beta_1 = 0, \quad H_1: \beta_1 \neq 0$$

为了检验假设 H_0 ，先分析对样本观察值 y_1, y_2, \dots, y_n 的差异，它可以用总的偏差平方和来度量，记为

$$S_{\text{总}} = \sum_{i=1}^n (y_i - \bar{y})^2$$

由正规方程组，有

$$\begin{aligned}
 S_{\text{总}} &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2
 \end{aligned}$$

令 $S_{\text{总}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, $S_{\text{剩}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, 则有

$$S_{\text{总}} = S_{\text{剩}} + S_{\text{回}}$$

上式称为总偏差平方和分解公式。 $S_{\text{回}}$ 称为回归平方和,它是由普通变量 x 的变化引起的,它的大小(在与误差相比下)反映了普通变量 x 的重要程度; $S_{\text{剩}}$ 称为剩余平方和,它是由试验误差,以及其他未加控制因素引起的,它的大小反映了试验误差及其他因素对试验结果的影响。关于 $S_{\text{回}}$ 和 $S_{\text{剩}}$,有下面的性质。

定理2:在线性模型假设下,当 H_0 成立时, $\hat{\beta}_1$ 与 $S_{\text{剩}}$ 相互独立,且 $S_{\text{剩}}/\sigma^2 \sim \chi^2(n-2)$, $S_{\text{回}}/\sigma^2 \sim \chi^2(1)$ 对 H_0 的检验有三种本质相同的检验方法,即七检验法;F检验法;相关系数检验法。

在介绍这些检验方法之前,先给出 $S_{\text{总}}$, $S_{\text{回}}$, $S_{\text{剩}}$ 的计算方法如下。

$$\begin{aligned}
 S_{\text{总}} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 \stackrel{\text{def}}{=} L_{yy} \\
 S_{\text{回}} &= \hat{\beta}_1^2 L_{xx} = \hat{\beta}_1 L_{xy} \\
 S_{\text{剩}} &= L_{yy} - \hat{\beta}_1 L_{xy}
 \end{aligned}$$

1. T 检验法

由定理1知 $(\hat{\beta}_1 - \beta_1)/(\sigma/\sqrt{L_{xx}}) \sim N(0,1)$,若令 $\hat{\sigma}^2 = S_{\text{剩}}/(n-2)$,则由定理2知, $\hat{\sigma}$ 为 σ^2 的无偏估计, $(n-2)\hat{\sigma}^2/\sigma^2 = S_{\text{剩}}/\sigma^2 \sim \chi^2(n-2)$,且 $(\hat{\beta}_1 - \beta_1)/(\sigma/\sqrt{L_{xx}})$ 与 $(n-2)\hat{\sigma}^2/\sigma^2$ 相互独立,故取检验统计量为

$$T = \frac{\hat{\beta}_1}{\hat{\sigma}} \sqrt{L_{xx}} \sim t(n-2)$$

由给定的显著性水平 α ,查表得 $t_{\alpha/2}(n-2)$,根据试验数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 计算 T 的值 t ,当 $|t| > t_{\alpha/2}(n-2)$ 时,拒绝 H_0 ,这时回归效应显著;当 $|t| \leq t_{\alpha/2}(n-2)$ 时,接受 H_0 ,此时回归效果不显著。

2. F 检验法

由定理2知,当 H_0 为真时,取统计量为

$$F = \frac{S_{\text{回}}}{S_{\text{剩}}(n-2)} \sim F(1, n-2)$$

由给定显著性水平 α , 查表得 $F_{\alpha}(1, n-2)$, 根据试验数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 计算 F 的值, 若 $F > F_{\alpha}(1, n-2)$ 时, 拒绝 H_0 , 表明回归效果显著; 若 $F \leq F_{\alpha}(1, n-2)$ 时, 接受 H_0 , 此时回归效果不显著。

3. 相关系数检验法

相关系数的大小可以表示两个随机变量线性关系的密切程度, 对于线性回归中的变量 x 与 Y , 其样本的相关系数为

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{\sqrt{L_{xy}}}{\sqrt{L_{xx}} \sqrt{L_{yy}}}$$

它反映了普通变量 x 与随机变量 Y 之间的线性相关程度, 故取检验统计量为

$$r = \frac{\sqrt{L_{xy}}}{\sqrt{L_{xx}} \sqrt{L_{yy}}}$$

对给定的显著性水平 α , 查相关系数表得 $r_{\alpha}(n)$, 根据试验数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 计算 R 的值, 当 $|r| > r_{\alpha}(n)$ 时, 拒绝 H_0 , 表明回归效果显著; 当 $|r| \leq r_{\alpha}(n)$ 时, 接受 H_0 , 表明回归效果不显著。

10.1.4 预测问题

在回归问题中, 若回归方程经检验效果显著, 这时回归值与实际值就拟合较好, 因而可以利用它对因变量 Y 的新观察值 y_0 进行点预测或区间预测。

对于给定的 x_0 , 由回归方程可得到回归值, 即

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

称 \hat{y}_0 为 y 在 x_0 的预测值, y 的测试值 y_0 与预测值 \hat{y}_0 之差称为预测误差。

在实际问题中, 预测的真正意义就是在一定的显著性水平 α 下, 寻找一个正数 $\delta(x_0)$, 使得实际观察值 y_0 以 $1-\alpha$ 的概率落入区间 $(\hat{y}_0 - \delta(x_0), \hat{y}_0 + \delta(x_0))$ 内, 即

$$P\{|Y_0 - \hat{y}_0| < \delta(x_0)\} = 1 - \alpha$$

由定理 1 知

$$Y_0 - \hat{y}_0 \sim N\left(0, \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}\right] \sigma^2\right)$$

又因 $Y_0 - \hat{y}_0$ 与 $\hat{\sigma}^2$ 相互独立, 且

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

所以

$$T = (Y_0 - \hat{y}_0) / \left[\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}} \right] \sim t(n-2)$$

故对给定的显著性水平 α , 求得 $\delta(x_0) = t_{\alpha/2}(n-1)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}}$, 故得 y_0 的置信水平为 $1 - \alpha$, 其预测区间为 $(\hat{y}_0 - \delta(x_0), \hat{y}_0 + \delta(x_0))$ 。

显而易见, y_0 的预测区间长度为 $2\delta(x_0)$, 对给定 α , x_0 越靠近样本均值 \bar{x} , $\delta(x_0)$ 越小, 预测区间长度小, 效果越好。当 n 很大, 并且 x_0 较接近 \bar{x} 时, 有

$$\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}} \approx 1, \quad t_{\alpha/2}(n-2) \approx u_{\alpha/2}$$

则预测区间近似为 $(\hat{y}_0 - u_{\alpha/2}\hat{\sigma}, \hat{y}_0 + u_{\alpha/2}\hat{\sigma})$ 。

10.1.5 SPSS 线性回归分析设置

选择菜单“分析 (Analyze) 回归 (Regression) 线性 (Linear)”, 则弹出如图 10-1 所示对话框, 此界面可以进行线性回归分析参数设置。

1. 变量选择设置

进行回归分析之前一定要对待分析的变量进行选择设置, 如图 10-1 所示, 其左边是对应的变量列表。

因变量 (Dependent) 选项: 选入因变量。

自变量 (Independent) 选项: 用于选入自变量, 可以选择一个或者多个自变量。

方法 (Method) 下拉菜单, 用于定义自变量进入模型的方法, 如图 10-2 所示, 各方法如下。

- 输入 (Enter);
- 步进 (Stepwise);
- 除去 (Remove);
- 后退 (Backward);
- 前进 (Forward)。



图 10-1 “线性 (Linear) 回归分析参数设置”对话框



图 10-2 方法 (Method) 设置

选择变量 (Selection Variable) 选项: 用于指定筛选变量, 只有满足条件的观测记录才会进入回归分析过程中, 选入变量后可以单击“规则 (Rule)”按钮, 则弹出如图 10-3 所示对话框。此对话框用于给定变量的筛选条件, 只有满足条件才可以进入回归分析。

个案标签 (Case Labels) 选项：用于选择变量作为每条记录的标签，通常选取记录号。

WLS 权重 (WLS Weight) 选项：选择权重变量进行加权最小二乘法回归分析，在分析时按照权重变量的大小给每条记录赋予不同的权重值。

2. 统计量 (Statistics) 设置

选择“统计量 (Statistics)”按钮，则弹出如图 10-4 所示对话框，此对话框主要进行计算统计量的设置，主要有如下几部分。



图 10-3 “规则 (Rule) 设置”对话框



图 10-4 “统计量 (Statistics) 设置”对话框

回归系数 (Regression Coefficient) 选项栏：此栏用于定义回归系数的输出情况，各选项具体作用如下所述。

- 估算值 (Estimates)：输出回归系数的估计值及其标准误、检验统计量，标准化的回归系数，为系统默认选择项。
- 置信区间 (Confidence Intervals)：输出每个回归系数的置信区间。
- 协方差矩阵 (Covariance Matrix)：输出每个自变量相关矩阵、方差、协方差矩阵。

模型拟合 (Model Fit)：选中后输出回归模型因变量列表、模型是否恰当的一些检验统计量，以及复相关系数 R 、决定系数 R^2 和调整的 R^2 、方差分析表等信息。

R 方变化量 (R squared Change)：输出模型拟合过程中 R^2 、 F 值和 P 值的改变情况。

描述 (Descriptives)：输出描述性统计量。

部分相关性和偏相关性 (Part and Partial Correlations)：输出自变量间的相关系数、部分相关系数和偏相关系数。

共线性诊断 (Colinearity Diagnostics)：输出多元线性回归中用于线性诊断的统计量。

残差 (Residuals) 选项栏：用于输出残差分析结果。

- 德宾-沃森 (Durbin-Watson)：输出 Durbin-Watson 残差序列相关性检验结果。
- 个案诊断 (Casewise Diagnostics)：输出超过规定的 n 倍标准差的残差列表或全部残差列表。

3. 图 (Plots) 设置

单击图 10-1 中的“图 (Plots)”按钮，则弹出如图 10-5 所示对话框，此对话框用于绘制各类图形。

变量列表：用来列举可以用来绘制图形的中间统计量，包括因变量（DEPENDNT）、标准化预测值（ZPRED）、标准化残差（ZRESID）、剔除残差（DRESID）、修正后的预测值（ADJPRED）、学生化的残差（SRESID）、学生化剔除残差（SDRESID）。

散点 1 的 1（Scatter 1 of 1）栏：从左侧候选变量框中选择变量进入 X、Y 轴选项框，定义需要绘制的回归分析诊断图或者预测图。

标准化残差图（Standardized Residual Plots）栏：选择绘制标准化残差图的类型，包括直方图（Histogram）、正态概率图（Normal Probability Plot）。

生成所有局部图（Produce all Partial Plots）：选择是否绘制每一个自变量与因变量残差的散点图。

4. 保存（Save）设置

单击图 10-1 中的“保存（Save）”按钮，则弹出如图 10-6 所示对话框，此界面主要用来设置存储分析的结果，对话框包括的选项如下。

预测值（Predicted Value）选项栏。

- 未标准化（Unstandardized）：保存模型对因变量的原始预测值。
- 标准化（Standardized）：保存标准化后的预测值，此预测值均值为 0，方差为 1。
- 调整（Adjusted）：保存去掉当前记录时，当前模型对该记录因变量的预测值。
- 平均值预测值的标准误差（S.E.of Mean Predictions）：保存预测值的标准差。

残差（Residuals）选项栏：用于保存回归诊断时所需的各种残差。

- 未标准化（Unstandardized）：保存模型对因变量的原始预测值。
- 标准化（Standardized）：保存使用 U 变换后的预测值，此预测值均值为 0，方差为 1。
- 学生化（Studentized）：保存学生化残差，即用 T 变换进行标准化后的残差。
- 删除后（Deleted）：保存删除当前记录后的残差。
- 学生化删除后（Studentized deleted）：保存删除当前记录后，用 T 变换进行标准化后的残差。

距离（Distances）选项栏：用于保存测量数据点离拟合模型距离的指标。

- 马氏距离（Mahalanobis）：保存记录值离样本平均值的距离。
- 库克（Cook's）距离：保存删除当前记录后，模型残差会发生的变化量。
- 杠杆值（Leverage Values）：测量该数据点的影响程度。

影响统计量（Influence Statistics）选项栏：保存用于判断强影响点的统计量。

- DfBeta（s）：保存去掉该记录值离样本平均值的距离。
- 标准化 DfBeta（s）（Standardized DfBeta（s））：保存标准化后的 DfBeta 值。
- DfFit：保存去掉该观测值点后观测值的变化值。
- 标准化 DfFit（Standardized DfFit）：保存标准化后的 DfFit 值。
- 协方差比率（Covariance ratio）：保存去掉该观测点之后协方差矩阵与含全部观察值的协方差矩阵的比率。

预测区间（Prediction Intervals）：选择是否给出均值和个体参考值的置信区间。

系数统计（Coefficient Statistics）：主要保存上述中间变量。

- 创建系数统计 (Create Coefficient Statistics) : 保存在一个新的数据文件中。
- 写入新数据文件 (Write a New Data File) : 直接保存在一个其他的文件中。

将模型信息输出到 XML 文件 (Export model information to XML file) : 将模型信息保存在一个 XML 文件之中, 单击“浏览 (Browse)”按钮选择路径。



图 10-5 “图 (Plots) 选项设置”对话框

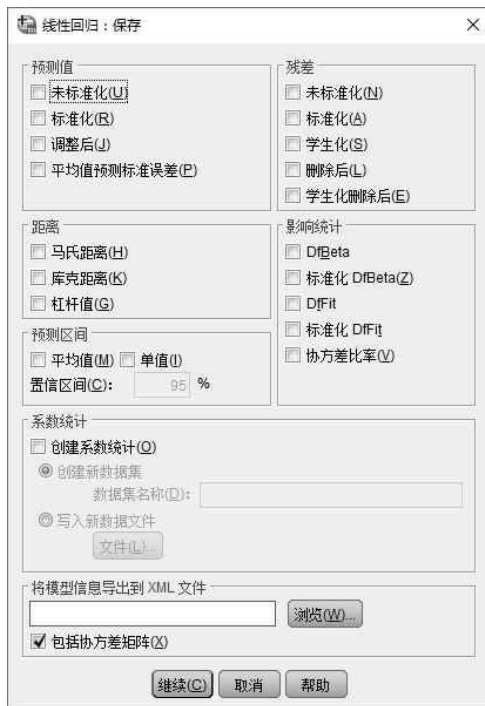


图 10-6 “保存 (Save) 设置”对话框

5. 选项 (Options) 设置

单击图 10-1 中的“选项 (Options)”按钮, 则弹出如图 10-7 所示对话框。此对话框用于设置回归分析的一些选项包括以下几项。

步进法条件 (Stepping Method Criteria) 选项栏: 用于设置变量纳入和排除标准。

在方程中包含常量 (Include Constant in Equation) 选项栏: 用于决定模型中是否包括常数项, 为默认选项。

缺失值 (Missing Values) 选项栏: 定义缺失值的处理方式。

- 成列排除个案 (Exclude Cases Listwise) : 只要数据中有变量值缺失就剔除该数据。
- 成对排除个案 (Exclude Cases Pariwise) : 仅当数据要分析的变量值缺失时才剔除该数据。
- 替换为平均值 (Replace with Mean) : 用变量均值代替变量缺失值。



图 10-7 “选项设置”对话框

10.1.6 回归分析模型的实例分析



结果文件

——附带光盘“PROGRAM\CH10\实例 10-1”文件夹



动画演示

——附带光盘“AVI\实例 10-1.avi”文件

本实例选择 SPSS 自带的数据集 polishing.sav，数据集包含 9 个变量，其数据集的格式如图 10-8 所示，下面对此数据集进行回归分析。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|-------|----|----|------|----|------------|----|---|----|----|----|
| 1 | bowl | 数字 | 1 | 0 | | 无 | 无 | 8 | 右 | 有序 | 输入 |
| 2 | case | 数字 | 1 | 0 | | 无 | 无 | 8 | 右 | 有序 | 输入 |
| 3 | dish | 数字 | 1 | 0 | | 无 | 无 | 8 | 右 | 有序 | 输入 |
| 4 | tray | 数字 | 1 | 0 | | 无 | 无 | 8 | 右 | 有序 | 输入 |
| 5 | plate | 数字 | 8 | 2 | | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 6 | type | 数字 | 8 | 2 | | {1.00, ... | 无 | 8 | 右 | 名义 | 输入 |
| 7 | diam | 数字 | 4 | 2 | | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 8 | time | 数字 | 6 | 2 | | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 9 | price | 数字 | 5 | 2 | | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 10-8 数据集 polishing.sav 的数据格式

1. 参数设置

首先进行回归分析之前，绘制数据的散点图，以观察数据走势。选择菜单“图形 (Graphs) 图表构建 (Chart Builder)”，弹出“图表构建器 (Chart Builder)”对话框。选中“散点/点图 (Scatter/Dot)”选项，并选择简单散点图 (Simple Scatter)。选中后则选择 time 变量为 Y 轴坐标，diam 变量为 X 轴坐标。然后单击“确定”按钮，则输出散点图形，如图 10-9 所示。

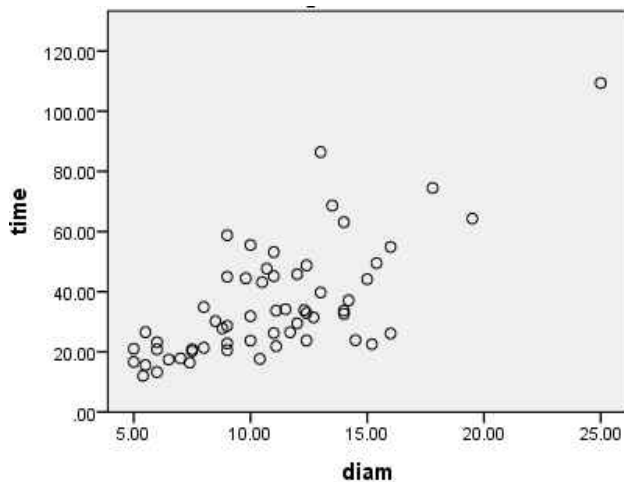


图 10-9 绘制散点图

然后选择菜单“分析 (Analyze) 回归 (Regression) 线性 (Linear)”，则弹出如图 10-10 所示对话框，选择变量 time 到“因变量 (Dependent)”变量框中，选择变量 diam 到“自变量 (Independent)(s)”变量框中，选择变量 type 到“个案标签 (Case Labels)”变量框中。



图 10-10 “线性 (Linear) 回归设置”对话框

然后单击图 10-10 中的“图 (Plots)”按钮，弹出如图 10-11 所示对话框，用于设置绘图形属性。选择变量*SDRESID 到 Y 变量框中，选择变量*ZPRED 到 X 变量框中。选中“直方图 (Histogram)”和“正态概率图 (Normal Probability Plot)”选项栏，然后单击“继续 (Continue)”按钮返回主界面。



图 10-11 “图 (Plots) 设置”对话框



图 10-12 “保存 (Save) 设置”对话框

单击“线性回归 (Linear Regression Dialog Box)”对话框中的“保存 (Save)”按钮，则弹出如图 10-12 所示对话框，选中“标准化 (Standardized)”选项栏，残差选项栏中的“标准化 (Standardized)”选项栏，以及库克距离 (Cook's) 和“杠杆值 (Leverage Values)”选项栏，然后单击“继续 (continue)”按钮返回主界面。

2. 结果分析

单击“线性回归 (Linear Regression Dialog Box)”对话框中的“确定 (OK)”按钮进行回归分析, 回归的相关参数分析结果如图 10-13 所示, 图中的结果可以得到以下式子。

$$\text{time} = 3.457 * \text{DIAM} - 1.955$$

| 系数 ^a | | | | | |
|-----------------|--------|--------|-------|-------|------|
| 模型 | 未标准化系数 | | 标准化系数 | t | 显著性 |
| | B | 标准误差 | Beta | | |
| 1 | (常量) | -1.955 | | -.362 | .719 |
| | diam | 3.457 | .467 | 7.407 | .000 |

a. 因变量: time

图 10-13 参数分析结果

方差分析结果如图 10-14 所示, 此表给出了模型的检验结果。F 值为 54.865, Sig 值为 0.000, 所以, 其显著性概率值远远小于 0.01, 所以, 显著的拒绝总体回归系数为 0 的假设。

| ANOVA ^a | | | | | | |
|--------------------|----|-----------|-----|-----------|--------|-------------------|
| 模型 | | 平方和 | 自由度 | 均方 | F | 显著性 |
| 1 | 回归 | 10287.173 | 1 | 10287.173 | 54.865 | .000 ^b |
| | 残差 | 10687.511 | 57 | 187.500 | | |
| | 总计 | 20974.684 | 58 | | | |

a. 因变量: time

b. 预测变量: (常量), diam

图 10-14 方差分析结果

图 10-15 所示的是直方图, 即是标准化残差的直方图, 且同时绘制了正态分布曲线, 可以看到残差基本符合正态分布。

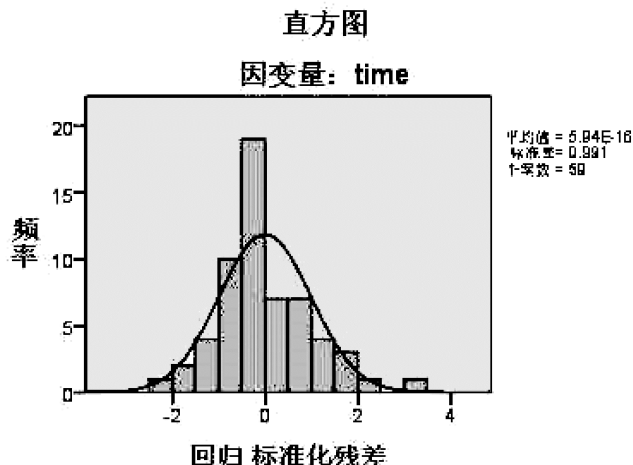


图 10-15 残差直方图

图 10-16 给出的是回归残差的散点图，图 10-17 为标准化残差对标准化预测值的散点图，从图中可以看到有异常点 Tray。

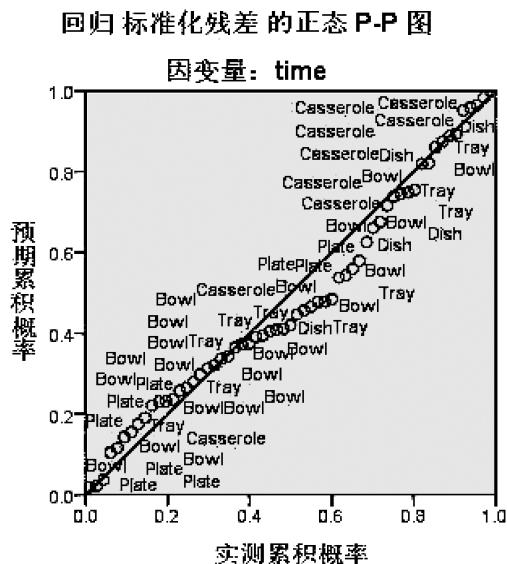


图 10-16 回归残差的散点图

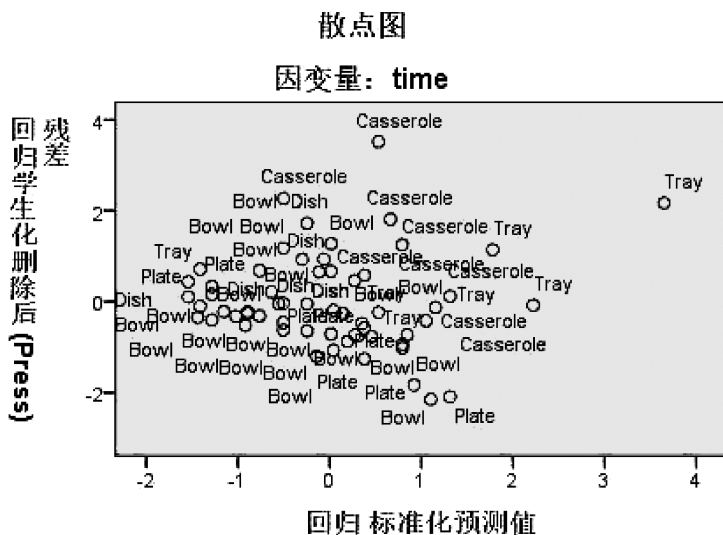


图 10-17 标准化残差对标准化预测值的散点图

10.2 非线性回归

如前所述，相对于线性回归来说，显然非线性回归分析过程要建立的是变量之间的非线性关系。当然非线性回归分析不像线性回归分析那样，非线性回归分析可以在变量之间任意建立某种非线性关系模型。

10.2.1 非线性回归分析的基本原理

对于很多非线性回归模型,由曲线回归的处理方法,同样可以经过简单函数的变换之后可以化为一元或者多元线性回归模型。但是在一般的情况之下,非线性模型难以精确的线性化,故需要给予特别的考虑。

一般的非线性回归模型可以表示为

$$Y = f(x, \beta) + \varepsilon$$

式中: x 是可以观察的独立随机变量; β 是待估的参数矢量; Y 是独立的观察变量,其平均数依赖于 x 和 β ; ε 是随机误差。函数形式 $f(\cdot)$ 是已知的。

非线性回归模型的解法主要有两个。一个是最小二乘法,即求矢量 Y 与集合 $f(x, \beta)$ 的最短距离为

$$\|Y - f(x, \beta)\| \xrightarrow{\beta} \min$$

另一种方法是极大似然法,即假设误差的分布密度函数 $g(x, \beta, \sigma^2)$ 已知,作为似然函数,再求其最大值即

$$L(\beta, \sigma^2) = \prod_{i=1}^n g(x, \beta, \sigma^2) \xrightarrow{\beta, \sigma^2} \max$$

当然,同处理曲线回归模型一样,处理非线性回归也可以通过变量变换,将非线性回归化为线性回归,然后用线性回归方法处理。

10.2.2 非线性回归参数设置

选择菜单“分析(Analyze) 回归(Regression) 非线性(Nonlinear)”,则弹出如图 10-18 所示对话框,此对话框用于设置非线性回归分析的各种参数。



图 10-18 “非线性(Nonlinear)回归参数设置”对话框

1. 变量选择设置

图 10-18 中的左上角为变量列表框,“因变量(Dependent)”选项框用于选入回归模型中的因变量。

2. 模型表达式(Model Expression)选项框

用于定义非线性回归模型的表达式。因为非线性模型实在是太多,SPSS 中直接提供了键盘和函数组(Function Group)选项栏来让用户自定义非线性模型的表达式。

3. 函数组(Function group)选项框

候选函数列表框,几乎涵盖了所有常用的函数类型。

4. 参数(Parameters)选项框

单击图 10-18 中的“参数(Parameters)”按钮,则弹出如图 10-19 所示对话框。各选项功能如下所述。

- 名称(Name):选择模型中的参数,参数名必须和 Model Expression 框中的参数名一致。
- 开始值(Starting Value):定义参数迭代初始值。
- 添加(Add) 更改(Change) 除去(Remove)按钮用于添加、改变和移出定义的参数迭代初值。
- 使用上一分析的开始值(Use starting values from previous analysis):在连续使用非线性回归模型时,是否以上次模型的参数拟合值作为本次模型的迭代初值。这样选择迭代初值,可以大大减少模型的迭代初次。

5. 保存(Save)设置

单击图 10-18 中的“保存(Save)”按钮,则弹出如图 10-20 所示对话框,该对话框用于定义需要保存的中间统计量,图 10-20 中各选项如下。

- 预测值(Predicted Values)。
- 残差(Residuals)。
- 导数(Derivatives)。
- 损失函数值。



图 10-19 “参数(Parameters)选项”对话框



图 10-20 “保存(Save)设置”对话框

6. 选项 (Options) 设置

单击图 10-18 中的“选项 (Options)”按钮,则弹出如图 10-21 所示对话框,此对话框用于设置参数迭代拟合过程中的一些选项。



图 10-21 “选项 (Options) 设置”对话框

标准误差的自助抽样估算 (Bootstrap Estimates of standard error): 选择是否利用 Bootstrap 方法估计参数的标准误。

估算方法 (Estimate Method): 定义参数的估计方法。

- 序列二次规划 (Sequential Quadratic Programming): 序列二次规划法。
 - 利文贝格—马夸特 (Levenberg-Marquardt) 方法: 只适用于无限制的模型。
- 序列二次规划 (Sequential Quadratic Programming): 定义序列二次规划法的迭代过程。
- 最大迭代次数 (Maximum Iterations): 定义最大迭代次数。
 - 步长限制 (Step Limit): 定义迭代过程中步长允许的最大变化值。
 - 最优性容差 (Optimality Tolerance)
 - 函数精度 (Function Precision): 方程精度, 即定义拟合的非线性回归模型的精度。
 - 无限步长 (Infinite Step Size): 定义迭代过程中所有参数允许的最大变化值。

利文贝格—马夸特 (Levenberg-Marquardt): 用于定义利文贝格—马夸特方法的迭代过程。

- 最大迭代次数 (Maximum Iterations): 定义最大迭代次数。
- 平方和收敛 (Sum-of-squares Convergence): 定义迭代终止条件。
- 参数收敛 (Parameter Convergence): 定义迭代停止条件。

7. 损失 (Loss) 设置

单击图 10-18 中的“损失 (Loss)”按钮,则弹出如图 10-22 所示对话框,此对话框用于定义回归模型的损失函数。

残差平方和 (Sum of Squared Residuals): 输出均方误差和损失函数。

用户定义的损失函数 (User-defined Loss Function): 用户自定义损失函数。

- 损失函数定义框: 用户定义的损失函数 (User-defined Loss Function) 选项下的列表框。
- 候选变量列表框: 即左上方的变量框, 列出了可疑用来定义损失函数的变量。

参数 (Parameters) 选项: 候选参数列表框。

函数组 (Function Group): 可选函数列表用于自定义函数, 也可用界面中的软键盘输入。



图 10-22 “损失 (Loss) 设置”对话框

8. 约束 (Constraints) 设置

单击图 10-18 中的“约束 (Constraints)”按钮, 则可以弹出如图 10-23 所示对话框。此对话框用于定义模型中迭代参数的限制条件。

- 未约束 (Unconstrained): 系统默认选项, 对参数不做任何限制。
- 定义参数约束 (Define Parameter Constraint): 用户自定义参数限制条件。其定义方法与前面的非线性回归模型的方法类似。



图 10-23 “约束 (Constraints) 设置”对话框

10.2.3 实例分析



结果文件

——附带光盘 “PROGRAM\CH10\实例 10-2” 文件夹



起始文件

——附带光盘 “AVI\实例 10-2.avi” 文件

本实例使用的数据集是 advert.sav，此数据集是关于某广告公司的数据，数据格式如图 10-24 所示。下面利用非线性回归分析来分析销售数量和广告投入的模型。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|--------|----|----|------|--------------------|---|----|---|----|----|----|
| 1 | advert | 数字 | 8 | 2 | Advertising spe... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 2 | sales | 数字 | 8 | 2 | Detrended sales | 无 | 无 | 8 | 右 | 标度 | 输入 |

At the bottom, there are tabs for '数据视图' (Data View) and '变量视图' (Variable View), with '变量视图' selected. The status bar at the bottom indicates 'IBM SPSS Statistics 处理程序就绪' and 'Unicode: ON'.

图 10-24 数据集 advert.sav 的格式

1. 参数设置

选择菜单“分析 (Analyze) 回归 (Regression) 非线性 (Nonlinear)”，则弹出如图 10-25 所示对话框，此对话框用于设置非线性回归分析的各种参数。选择变量 sales 到“因变量 (Dependent)”选项栏中，在“模型表达式 (Model Expression)”选项栏中输入模型表达式： $b1+b2*\exp(b3*advert)$ ，然后单击“参数 (Parameters)”按钮，把初始值 $b1=13$ 、 $b2=-6$ 、 $b3=-1.33$ 填入“参数 (Parameters)”选项栏中。



图 10-25 “非线性 (Nonlinear) 回归参数设置”对话框

然后单击“约束 (Constraints)”按钮则弹出如图 10-26 所示对话框，设置约束条件， $b1$ 大于等于 0， $b2$ 和 $b3$ 小于等于 0。然后单击“继续 (Continue)”按钮返回主界面。

单击“保存 (Save)”按钮，弹出如图 10-27 所示对话框，选中“预测值 (Predicted Values)”和“残差值 (Residuals)”选项栏，然后单击“继续”按钮返回主界面。



图 10-26 “约束 (Constraints) 设置”对话框



图 10-27 “保存 (Save) 设置”对话框

2. 结果分析

单击图 10-25 中的“确定”按钮进行非线性回归分析，如图 10-28 所示的是参数估计结果，从结果可以看到非线性模型方程如下。

$$y = 12.904 - 11.268 \exp(-0.496x)$$

图 10-28 的右边是置信区间为 95% 时的估计值的上限和下限。

| 参数估计值 | | | | |
|-------|---------|-------|----------|--------|
| 参数 | 估算 | 标准误差 | 95% 置信区间 | |
| | | | 下限 | 上限 |
| b1 | 12.904 | .610 | 11.636 | 14.173 |
| b2 | -11.268 | 1.581 | -14.556 | -7.979 |
| b3 | -.496 | .138 | -.782 | -.209 |

图 10-28 参数估计结果

图 10-29 是关于估计值的相关性结果，图 10-30 是方差分析结果，从结果可以看到 R 方等于 0.909 大于 0.90，说明拟合模型能解释因变量大于 90% 的变异，所以拟合结果较好。

| 参数估计值相关性 | | | |
|----------|-------|-------|-------|
| | b1 | b2 | b3 |
| b1 | 1.000 | .693 | .946 |
| b2 | .693 | 1.000 | .871 |
| b3 | .946 | .871 | 1.000 |

图 10-29 参数估计的相关性结果

| ANOVA ^a | | | |
|--|----------|-----|---------|
| 源 | 平方和 | 自由度 | 均方 |
| 回归 | 2748.519 | 3 | 916.173 |
| 残差 | 6.778 | 21 | .323 |
| 修正前总计 | 2755.297 | 24 | |
| 修正后总计 | 74.520 | 23 | |
| 因变量: Detrended sales | | | |
| a. R 方 = 1 - (残差平方和) / (修正平方和) = .909. | | | |

图 10-30 方差分析结果

10.3 Logistic 回归

Logistic 回归又称逻辑回归分析，主要在流行病学中应用较多，比较常用的情形是探索某疾病的危险因素，根据危险因素预测某疾病发生的概率等。例如，想探讨胃癌发生的危

危险因素,可以选择两组人群,一组是胃癌组,另一组是非胃癌组,两组人群肯定有不同的体征和生活方式等。这里的因变量就是是否胃癌,即“是”或“否”,为两分类变量,自变量就可以包括很多,例如,年龄、性别、饮食习惯、幽门螺杆菌感染等。自变量既可以是连续的,也可以是分类的。通过 Logistic 回归分析,就可以大致了解到底哪些因素是胃癌的危险因素。

Logistic 回归与多重线性回归实际上有很多相同之处,最大的区别就在于它们的因变量不同,其他的基本都差不多,正是因为如此,这两种回归可以归于同一个家族,即广义线性模型(Generalized Linear Model)。这一家族中的模型形式基本上都差不多,不同的就是因变量不同,如果是连续的,就是多重线性回归,如果是二项分布,就是 Logistic 回归,如果是 Poisson 分布,就是 Poisson 回归,如果是负二项分布,就是负二项回归等。只要注意区分它们的因变量就可以了。总之,Logistic 回归是多元线性回归的延伸。

10.3.1 Logistic 回归模型概述

Logistic 回归的因变量可以是二分类的,也可以是多分类的,但是二分类的更为常用,也更加容易解释。所以,实际中最为常用的就是二分类的 Logistic 回归。

Logistic 回归的主要用途:一是寻找危险因素,正如上面所说的寻找某一疾病的危险因素等;二是预测,如果已经建立了 Logistic 回归模型,则可以根据模型,预测在不同的自变量情况下,发生某病或某种情况的概率有多大。三是判别,实际上跟预测有些类似,也是根据 Logistic 模型,判断某人属于某病或属于某种情况的概率有多大,也就是看一下这个人有多大的可能性是属于某病。

上述是 Logistic 回归最常用的三个用途,实际中的 Logistic 回归用途是极为广泛的,Logistic 回归几乎已经成了流行病学和医学中最常用的分析方法,因为它与多重线性回归相比有很多的优势,这些优势将在以后的文章中一一介绍。本篇文章主要是先让大家对 Logistic 回归有一个初步的了解,以后会对该方法进行详细的阐述。

1. logistic 函数

logistic 函数又称为增长函数,是由美国科学家 Robert.B.Pearl 和 Lowell.J.Reed 在研究果蝇的繁殖中提出来的,其一般的表达式为

$$p = \frac{1}{1 + \exp(-z)}, \quad -\infty < z < +\infty$$

由于 $1-p = \frac{1}{1 + \exp(z)}$, 所以有 $\frac{p}{1-p} = \frac{1 + \exp(z)}{1 + \exp(-z)} = \exp(z)$, 两边取对数可以得到

$$\ln\left(\frac{p}{1-p}\right) = z$$

2. Logistic 回归模型

首先令 Y 服从二项分布,取值为 0, 1。 $Y=1$ 的概率为 $\pi(Y=1)$, 则 m 个自变量分别为 X_1, X_2, \dots, X_m 所对应的 Logistic 回归模型为

$$\pi(Y=1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)}$$

或者写为

$$\text{logistic}[\pi(Y=1)] = \ln \frac{\pi(Y=1)}{1 - \pi(Y=1)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

式中, β_0 为截距, β_i 为 X_i 对应的偏回归系数。

令 $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$, 由上面所述, 即

$$L = \ln\left(\frac{p}{1-p}\right)$$

上式称为对数单位, $\frac{p}{1-p}$ 称为机会比率, 即有利于出现某一状态的机会大小。令

$$L = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

则上式模型也为 Logistic 回归模型。

与一般的线性概率模型相比, Logistic 回归模型有以下几个优点:

随着 p 从 0 到 1, L 从负无穷大到正无穷大, 即虽然概率受到 0~1 的限制, 但是 L 却不受限制。

- 虽然概率 p 与各个自变量之间是非线性的, 但是 L 与各个自变量之间是线性的。
- Logistic 回归模型系数的经济意义是 X 每变化一个单位, 有利机会对数变化的程度。

10.3.2 二元 Logistic 回归模型参数设置

选择菜单“分析 (Analyze) 回归 (Regression) 二元 Logistic (Binary Logistic)”, 则弹出如图 10-31 所示对话框。此对话框用于设置二元 Logistic (Binary Logistic) 模型的各种参数, 此对话框有以下几个部分。

1. 变量设置

图 10-31 左边为变量列表, “因变量 (Dependent)” 选项栏为选入 Logistic 回归的因变量, 只可以选择一个二值变量, 否则最后的输出结果会出现警告。“协变量 (Covariates)” 选项栏用于选择 Logistic 回归的自变量, 当候选变量框中同时选中两个以下或者两个以上变量时, 激活 “>a*b>” 按钮。

选择 “变量 (Selection Variable)” 选项框用于选入筛选变量, 当选入变量后则会激活其后的 “规则 (Rule)” 按钮, 单击此按钮, 则弹出如图 10-32 所示对话框, 此对话框主要用于给定变量的筛选条件。

2. 方法 (Method) 下拉菜单

用于选择变量进入模型的方法, 如图 10-33 所示。

- 输入 (Enter): 强行进入法。
- 向前 (Forward): 条件 (Conditional) / LR/Ward: 依据条件参数似然比检验结果/偏似然比检验结果/Ward 检验结果剔除变量的向前剔除方法。

- 向后 (Backward) : 条件 (Conditional) /LR/Ward : 依据条件参数似然比检验结果/偏似然比检验结果/Ward 检验结果剔除变量的向后剔除方法。

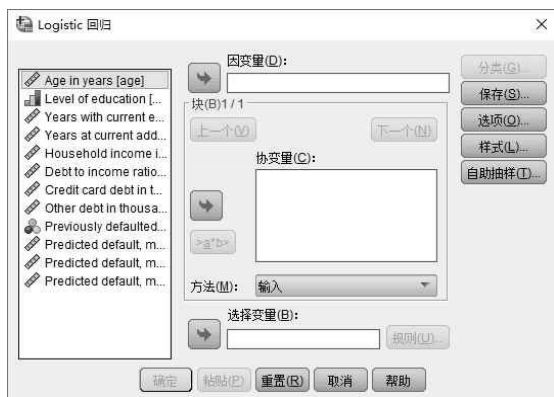


图 10-31 “二元 Logistic (Binary Logistic) 模型设置”对话框



图 10-32 “筛选变量设置”对话框

3. 分类 (Categorical) 选项

单击图 10-31 中的“分类 (Categorical)”按钮,则弹出如图 10-34 所示的“分类 (Categorical)”对话框,此对话框主要用于将某些数值型自变量定义为分类变量。

协变量 (Covariates) 选项栏: 用于存放候选变量列表框。

分类协变量 (Categorical Covariates) 选项框: 选择要将其定义为分类变量的变量。



图 10-33 “方法 (Method) 设置”下拉菜单



图 10-34 “分类 (Categorical) 选项设置”对话框

更改对比 (Change Contrast) 选项栏: 用于设置每个变量的哑变量组中的具体取值和对照组。其中的对比 (Contrast) 下拉菜单用于选择哑变量取值; 参考类别 (Reference Category) 选项栏用于设置选择第一个或者最后一个水平为对照。

4. 保存 (Save) 设置

单击图 10-31 中的“保存 (Save)”按钮,则弹出如图 10-35 所示对话框,此对话框用于定义需要保存的中间统计量,各个选项功能如下所述。

预测值 (Predicted Values) 选项栏: 此栏用于设置要保存的预测值。

- 概率 (Probabilities) : 保存每个观测值的预测概率值。
- 组成员 (Group Membership) : 保存根据预测概率值判断观测值所属的类别。
影响 (Influence) 选项栏 : 定义用于判断强影响点的统计量。
- 库克 (Cook's) 距离 : 保存删除当前记录后, 模型残差会发生的变化量。
- 杠杆值 (Leverage Values) : 保存杠杆值, 即测量该数据点的影响强度。
- DfBeta (D) : 保存去掉该观测值后回归系数的变化值。
残差 (Residuals) 选项栏 : 用于保存各种残差。
- 未标准化 (Unstandardized) : 保存模型预测值对因变量观测值的原始残差。
- 分对数 (Logit) : 保存分对数残差。
- 学生化 (Studentized) : 保存学生化残差。
- 标准化 (Standardized) : 保存用 U 变换进行标准化后的残差, 此时均值为 0, 标准差为 1。
- 偏差 (Deviance) : 保存 Deviance 残差。

将模型信息输出到 XML 文件 (Export model information to XML file) : 将模型信息导入到 XML 文件之中。

包含协方差矩阵 (Include the covariance matrix) : 保存变量间的相关矩阵。

5. 选项 (Options) 设置

单击图 10-31 中的“选项 (Options)”按钮, 则弹出如图 10-36 所示对话框, 此对话框主要由如下几个部分组成。

统计和图 (Statistics and Plots) 选项栏 : 用于定义一些重要的统计量和统计图形。

- 分类图 (Classification Plots) : 绘制因变量实际分类和预测分类关系图。
- 霍斯默—莱梅肖拟合优度 (Hosmer-Lemeshow goodness-of-fit) : 计算霍斯默—莱梅肖拟合优度指标。
- 个案的残差列表 (Casewise listing of residuals) : 对于记录逐条列出或满足一定条件列出其残差和概率预测值、预测分类, 以及实际分类。
- 估计值的相关性 (Correlations of Estimates) : 计算参数估计值的相关系数矩阵。
- 迭代历史记录 (Iteration history) : 列出极大似然估计每一步的迭代估计值。
- Exp 的置信区间 (B) : 计算参数值 95% 的置信区间。

显示 (Display) 选项栏 : 定义是否输出模型迭代过程中每一步的结果。

- 在每个步骤中 (At each step) : 输出每一步的结果。
- 在最后一个步骤中 (At last step) : 输出最后一步的结果。

步进概率 (Probability for Stepwise) : 定义模型中变量进入或者移出的概率值。进入 (Entry) 默认为 0.05 ; 除去 (Removal) 默认为 0.10。

分类分界值 (Classification cutoff) : 定义预测观测值分类的概率值大小。

最大迭代次数 (Maximun Iterations) : 最大迭代次数。

在模型中包含常量 (Include constant in model) : 定义模型中是否包含常数项。



图 10-35 “保存 (Save) 设置”对话框



图 10-36 “选项 (Options) 设置”对话框

10.3.3 实例分析



结果文件——附带光盘“PROGRAM\CH10\实例 10-3”文件夹



起始文件——附带光盘“AVI\实例 10-3.avi”文件

本实例分析 SPSS 自带的数据集 bankloan.sav，此数据集为银行贷款的用户信用记录数据，下面利用 Binary Logistic 回归分析来研究用户信用风险。数据集 bankloan.sav 的数据格式如图 10-37 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|----|----------|----|----|------|---------------------|------------------|----|----|----|----|----|
| 1 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 4 | 右 | 标度 | 输入 |
| 2 | ed | 数字 | 4 | 0 | Level of education | {1, Did not c... | 无 | 4 | 右 | 有序 | 输入 |
| 3 | employ | 数字 | 4 | 0 | Years with curr... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 4 | address | 数字 | 4 | 0 | Years at curren... | 无 | 无 | 7 | 右 | 标度 | 输入 |
| 5 | income | 数字 | 8 | 2 | Household inco... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 6 | debtinc | 数字 | 8 | 2 | Debt to income... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 7 | creddebt | 数字 | 8 | 2 | Credit card deb... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 8 | othdebt | 数字 | 8 | 2 | Other debt in th... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 9 | default | 数字 | 4 | 0 | Previously defa... | {0, No}... | 无 | 7 | 右 | 名义 | 输入 |
| 10 | preddef1 | 数字 | 11 | 5 | Predicted defau... | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 11 | preddef2 | 数字 | 11 | 5 | Predicted defau... | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 12 | preddef3 | 数字 | 11 | 5 | Predicted defau... | 无 | 无 | 11 | 右 | 标度 | 输入 |

图 10-37 数据集 bankloan.sav 的格式

1. 参数设置

首先设置随机数产生器，选择菜单“转换 (Transform) 随机数生成器 (Random Number Generators)”，弹出如图 10-38 所示对话框，选中“设置起点 (Set Starting Point)”选项栏，并在其下的“固定值 (Fixed Value)”选项栏的值 (Value) 中输入 9191972，然后单击“确定”按钮。

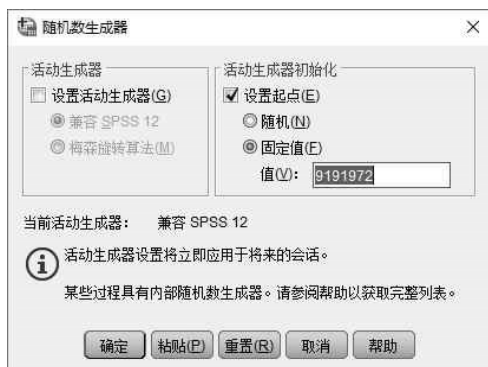


图 10-38 “随机数生成器 (Random Number Generators) 设置”对话框

然后再选中菜单“转换 (Transform) 计算变量 (Compute Variable)”，弹出如图 10-39 所示对话框，用于设置随机数的分布类型。输入 validate 到“目标变量 (Target Variable)”选项栏中，输入 RV.BERNOULLI (0.7) 到“数值表达式 (Numeric Expression)”选项栏中，表示筛选变量 validate 的取值服从参数为 0.7 的 bernoulli 分布。



图 10-39 “计算变量 (Compute Variable) 设置”对话框

然后单击图 10-39 中的“如果 (If)”按钮，弹出如图 10-40 所示对话框。选择“如果个案满足条件则包括 (Include if case satisfies condition)”选项栏，输入 MISSING(default) = 0 到其下的变量框中，然后单击“继续 (Continue)”按钮返回主界面。

接着单击图 10-39 中的“确定”按钮，设置完成，生成变量 validate。下面进行二项 Logistic 分析的设置阶段。选择菜单“分析 (Analyze) 回归 (Regression) 二元 Logistic (Binary Logistic)”，弹出如图 10-41 所示对话框。选择变量 Previously defaulted 到“因变量”选项栏中，选择变量 age、ed、employ、address、income、debtinc、creddebt，以及 othdebt 到“协变量”选项栏中，选择变量 validate 到“选择变量 (Selection Variable)”选项栏中。



图 10-40 “如果 (If) 设置”对话框

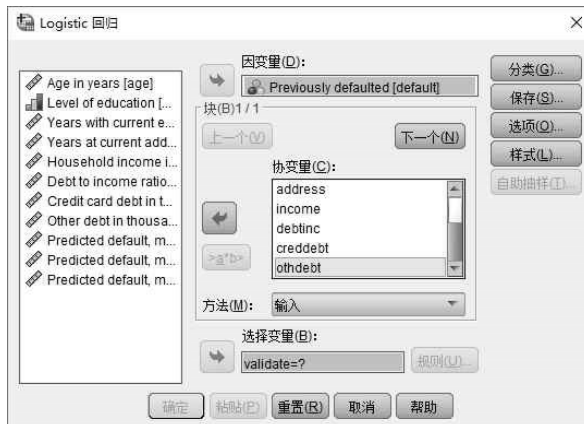


图 10-41 二项 Logistic “Binary Logistic 设置”对话框

然后单击图 10-41 中的“规则 (Rule)”按钮，弹出如图 10-42 所示对话框，选项栏中填入 1，即只选择那些 validate 等于 1 的变量进行分析，并单击“继续 (Continue)”按钮返回主界面。

然后单击图 10-41 中的“分类 (Categorical)”按钮，弹出如图 10-43 所示对话框。选择变量 Level of education 到“分类协变量 (Categorical Covariate)”选项栏中，并单击“继续 (Continue)”按钮返回主界面。



图 10-42 “规则 (Rule) 设置”对话框

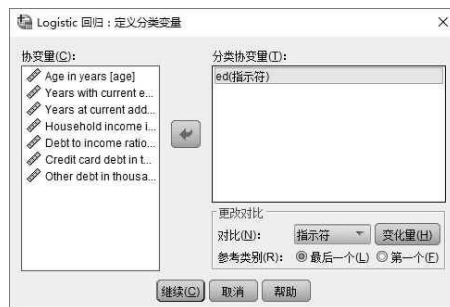


图 10-43 “分类 (Categorical) 设置”对话框

紧接着单击图 10-41 中的“保存 (Save)”按钮，弹出如图 10-44 所示对话框。选择“概率 (Probabilities)”、“Cook 距离”、“学生化 (Studentized)”选项栏，并单击“继续 (Continue)”按钮返回主界面。

继续单击图 10-41 中的“选项 (Options)”按钮，弹出如图 10-45 所示对话框。选择“分类图 (Classification Plots)”、“霍斯默—莱梅肖拟合拟合度 (Hosmer-Lemeshow goodness-of-fit)”选项栏，并单击“继续 (Continue)”按钮返回主界面。



图 10-44 “保存 (Save)”设置对话框



图 10-45 “选项 (Options)”设置对话框

2. 结果分析

设置完成以后单击二元 Logistic (Binary Logistic) 主界面中的“确定 (OK)”按钮，则进行分析，结果如下。首先是个案处理结果，如图 10-46 所示。给出了基本的统计信息，共有 499 个，占总体 58.7% 的观测记录应用于模型的分析。图 10-47 给出的是教育水平的分类变量编码。

下面是模型的基本信息，如图 10-48 所示。

图 10-49 是霍斯默—莱梅肖拟合检验结果，显著性 0.502，刚刚大于 0.5，说明模型的拟合结果不是很好。

| 个案处理摘要 | | |
|---------------------|-----|-------|
| 未加权个案数 ^a | 个案数 | 百分比 |
| 选定的个案 | 499 | 58.7 |
| 包括在分析中的个案数 | | |
| 缺失个案数 | 0 | .0 |
| 总计 | 499 | 58.7 |
| 未选定的个案 | 351 | 41.3 |
| 总计 | 850 | 100.0 |

a. 如果权重处于生效状态，请参阅分类表以了解个案总数。

图 10-46 个案处理结果

| 分类变量编码 | | | | | | |
|--------------------|------------------------------|-----|-------|-------|-------|-------|
| | | 频率 | 参数编码 | | | |
| | | | (1) | (2) | (3) | (4) |
| Level of education | Did not complete high school | 266 | 1.000 | .000 | .000 | .000 |
| | High school degree | 134 | .000 | 1.000 | .000 | .000 |
| | Some college | 69 | .000 | .000 | 1.000 | .000 |
| | College degree | 25 | .000 | .000 | .000 | 1.000 |
| | Post-undergraduate degree | 5 | .000 | .000 | .000 | .000 |

图 10-47 教育水平的分类变量编码

| 模型摘要 | | | |
|------|----------------------|----------------|----------|
| 步骤 | -2 对数似然 | 考克斯-斯奈尔 R 方 | 内戈尔科 R 方 |
| 1 | 388.480 ^a | .290 | .431 |

a. 由于参数估算值的变化不足 .001，因此估算在第 6 次迭代时终止。

图 10-48 模型信息

| 霍斯默-莱梅肖检验 | | | |
|-----------|-------|-----|------|
| 步骤 | 卡方 | 自由度 | 显著性 |
| 1 | 7.324 | 8 | .502 |

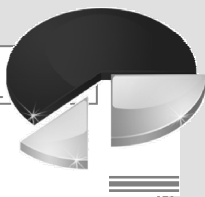
图 10-49 Hosmer-Lemeshow 检验结果

图 10-50 是预测与分类的结果，给出了关于观测值和预测值的列联表，对于模型，建模所有的 124 个拖欠用户中有 59 个判断正确，正确率是 47.6%。建模所有的 375 个无拖欠用户中有 350 个判断正确，所以正确率为 93.3%，对于数据总的回判正确率为 82%。

| 分类表 ^a | | | | | | | |
|------------------|----------------------|----------------------------|-----|-------|----------------------------|-----|-------|
| 实测 | | 预测 | | | | | |
| | | 选定的个案 ^b | | | 未选定的个案 ^{c,d} | | |
| | | Previously defaulted No | Yes | 正确百分比 | Previously defaulted No | Yes | 正确百分比 |
| 步骤 1 | Previously defaulted | No | 350 | 25 | 93.3 | 129 | 13 |
| | Yes | 65 | 59 | 47.6 | 27 | 32 | 54.2 |
| 总体百分比 | | | | 82.0 | | | 80.1 |

a. 分界值为 .500
b. 选定的个案 validate EQ 1
c. 未选定的个案 validate NE 1
d. 由于自变量中缺少值或者分类变量的值超出选定个案的范围，因此未对某些未选定的个案进行分类。

图 10-50 预测分类结果



第 11 章 相关分析

相关分析 (Correlation Analysis) 是研究现象之间是否存在某种依存关系, 并对具有依存关系的现象探讨其相关方向, 以及相关程度, 是研究随机变量之间的相关关系的一种统计方法。相关关系是一种非确定性的关系, 例如, 以 X 和 Y 分别记一个人的身高和体重, 或分别记每公顷施肥量与每公顷小麦产量, 则 X 与 Y 显然有关系, 而又没有确切到可由其中的一个去精确地决定另一个的程度, 这就是相关关系。

本章将利用 SPSS 来实现相关分析。



本讲内容

- 相关分析概述
- 双变量过程
- Partial 过程
- Distances (距离) 过程

11.1 相关分析概述

统计分析的一项重要课题是, 根据辩证唯物主义和历史唯物主义关于事物普遍联系和相互作用的原理来进行社会经济现象相互联系的分析研究。

可以列举许多关于社会经济生活相互依存、相互制约、相互影响的例子。例如, 企业规模和经营费用的关系、工资增长和劳动生产率的关系、家庭收入水平和支出的关系、劳动机械化水平与劳动生产率的关系等。无疑, 从数量上研究这些现象相互依存关系, 分析现象变动的影响因素和作用强度, 对于加强经济的科学管理, 发挥统计工作的职能有其现实意义。

相关分析就是研究两个或两个以上变量之间相互关系的统计分析方法, 它是研究二元总体和多元总体的重要方法。

11.1.1 相关关系

进行相关分析, 首先必须明确什么是相关关系。对社会生活中各种现象所作的统计研究, 要做到数量上反映现象间复杂的相互关系, 首先要凭借研究者所掌握的科学知识、工作

能力和判断能力做定性分析, 以免把不相关或虚假相关现象拿来进行分析、定性分析, 对总体一系列标志找到其中有联系的成对标志, 确定哪个是因素标志, 哪个是结果标志, 即自变量和因变量。因果关系是客观世界普遍联系和相互制约的重要表现形式。相关分析就是对总体中确实具有联系的标志进行分析, 其主体是对主体中具有因果关系的标志的分析。

因素标志是决定结果标志发展的条件, 根据结果标志对因素标志的不同反应, 可以把现象总体数量上所存在的依存关系划分为两种不同的类型, 一种是函数关系; 另一种是相关关系。函数关系是当因素标志的数量确定后, 结果标志的数量也随之完全确定。函数关系在自然科学中常常遇到。社会经济现象也会遇到函数关系, 例如, 在计价工资制的情况下, 工资总额与工人加工的零件数量成函数关系; 商品销售额与销售量之间成函数关系。函数关系可以用 $y = f(x)$ 来表示, 它表明数量之间联系的一种形式。

相关关系是不完全确定的随机关系。在相关关系的情况下, 因素标志的每一个数值, 都有可能若有若干个结果标志的数值。所以, 相关关系是一种不完全的依存关系。例如, 工人技术水平的提高, 使得劳动生产率提高, 但不意味着做同样工作的几个同级工人都有同样高的劳动生产率。又如, 水稻播种的株行距确定了, 但每亩的产量确有多有少, 并不随株行距完全确定。究其原因是现象在数量上受各种各样因素的影响, 其中错综复杂的关系有些属于人们暂时还没有认识到的, 有些虽已被认识但无法控制, 而计量上的可能误差, 都造成现象之间变量关系的不确定性。但是不确定的变量关系还是有规律可循的, 经过大量观察, 会发现许多现象变量之间确实存在着某种规律性, 这就是由于大多数法则的作用, 把那些影响结果标志数值的其他一些次要、偶然因素抵消、抽象了, 使相关关系通过平均值明显地表现出来。

函数关系与相关关系的联系表现在, 对具有相关关系的现象分析时, 必须利用相应的函数关系数学表达式, 来表明现象之间的相关方程式。相关关系是相关分析的研究对象, 函数关系是相关分析的工具。也可以说函数关系是相关关系的特殊表现形式。

现象的相关关系可以按不同标志加以区分如下。

- 按相关程度分为完全相关、不完全相关和不相关。
- 按相关的方向分为正相关和负相关。
- 按相关的形式分为线性相关和曲线相关。
- 按相关的影响因素分为单相关和复相关。

统计实践中, 经常分析若干个因素标志对结果标志的影响, 即为复相关, 也就是多元相关。

11.1.2 相关图形和相关系数

怎样确定变量之间的相关关系呢? 一般可以通过图形观测和指标测量来确定相关关系。图形观测方法可以通过变量之间的三点图来进行分析确定。指标测量观测是通过计算相关系数来确定的。首先, 介绍简单的图形观测方法。

1. 相关图形

统计分析中的相关图, 可以直观地判断现象之间大致呈现何种关系的形式。相关图是相关分析的重要方法。

对现象总体两种相关标志作相关分析, 研究其相互依存关系, 首先要通过对实际调查取得一系列成对的标志值资料, 作为相关分析原始数据。

利用笛卡儿坐标系第一象限, 把自变量置于横轴之上, 因变量置于纵轴之上, 而将两因变量对应的变量值用坐标点形式描绘出来, 用以表明相关点分布状况的图形, 就是相关图。

2. 相关系数

相关图只是大体上反映现象的相关程度。因此, 还应该利用科学的方法进一步分析相关的密切程度, 统计上的相关系数就是用来说明相关关系密切程度的统计指标。

相关关系的特点如下:

- 两个变量是对等的, 不分自变量与因变量。相关系数只有一个。
- 相关系数有正负号, 反映正相关与负相关。

相关关系通常用小写字母 r 来表示。现在假设有两个变量 x 和 y , 根据样本数据计算相关系数的方法, 是利用积差法来计算相关系数, 计算公式为

$$r = \frac{\frac{1}{n} \sum (x - \bar{x})(y - \bar{y})}{\sqrt{\frac{\sum (x - \bar{x})^2}{n}} \sqrt{\frac{\sum (y - \bar{y})^2}{n}}}$$

其中, 分子是两变量的协方差, 分母是两变量的标准差。所以

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

以上公式简化, 得

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}} = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}}$$

将上式展开, 得

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

从以上公式中可以看出, r 的符号取正还是取负只决定于分子 L_{xy} 值是正还是负, r 的符号与 L_{xy} 的符号保持一致。

相关系数 r 的符号反映相关关系的方向, 其绝对值的大小则反映变量相关关系的密切程度。相关系数的绝对值不会超过 1, 所以 r 的取值范围为

$$0 \leq |r| \leq 1$$

r 的值不同, 散点图的形状便随之不同, 它们反映着两变量相关的方向、程度上的差异, 具体情况如下。

- $r = 1$: 表示完全正线性相关。
- $r > 0$: 表示正线性相关。
- $r = 0$: 表示不存在线性关系。
- $r < 0$: 表示负线性相关。
- $r = -1$: 表示完全负线性相关。

这里要注意的是,当 $r=0$ 时,虽然不存在线性关系,但是变量之间也可能存在其他的关系。

11.1.3 SPSS 的相关分析功能简介

SPSS 中的相关分析可以通过选择菜单“分析 (Analyze) 相关 (Correlate)”来实现,如图 11-1 所示。



图 11-1 “相关分析 (Correlate)” 菜单

相关 (Correlate) 子菜单主要包括如下的几个过程。

- 双变量 (Bivariate): 两变量相关分析。包括两个连续变量之间的相关和两个等级变量之间的秩相关。
- 偏相关 (Partial): 偏相关分析。当两变量的取值受其他变量的影响,则采用偏相关分析的方法控制其他变量的影响,研究两变量间的相关分析。
- 距离 (Distances): 距离分析。主要用于分析同一变量内观测值之间或者多个变量之间的相似或者不相似程度。
- 典型相关性 (Canonical correlation): 用以分析两组变量间关系的一种方法。

11.2 双变量 (Bivariate) 过程

两样本相关分析即是研究两个变量之间相关的统计方法。相关系数主要有三种: Pearson、Spearman (斯皮尔曼)、Kendall 相关系数。

11.2.1 双变量相关分析简介

1. Pearson 相关系数

用于对定距变量的数据进行计算,即分析两个连续性数据之间的关系,其计算公式为

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}$$

上式只是代表了样本的相关系数。

2. Spearman 相关系数

Spearman 等级相关系数 (Spearman's Coefficient of Rank Correlation) 是历史上最早 (1904 年) 测定两个样本相关强度的重要指标, 记为

$$r_s = \frac{\text{cov}(R_x, R_y)}{S_{R_x} \cdot S_{R_y}} = \frac{\sum_{i=1}^n (R_{xi} - \bar{R}_x)(R_{yi} - \bar{R}_y)}{\sqrt{\sum_{i=1}^n (R_{xi} - \bar{R}_x)^2 \sum_{i=1}^n (R_{yi} - \bar{R}_y)^2}}$$

显然

$$\bar{R}_x = \bar{R}_y = \frac{n+1}{2}, \text{ 记: 等级差 } d_i = R_{xi} - R_{yi}, \text{ 则}$$

$$r_s = 1 - \frac{6 \times \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

式中, n 为变量的对子数, d 为秩次之差。

当相同秩次较多时, 会影响 $\sum d^2$ 值, 应采用下式计算校正的等级相关系数, 即

$$r'_s = \frac{\frac{n^3 - 3}{6} - (t_x + t_y) - \sum d^2}{\sqrt{\left(\frac{n^3 - n}{6} - 2t_x\right)\left(\frac{n^3 - n}{6} - 2t_y\right)}}$$

式中, t_x 、 t_y 的计算公式相同, 均为 $\sum \frac{t_i^3 - t_i}{12}$ 。

在计算 t_x 时, t_i 为 x 变量的相同秩次数; 在计算 t_y 时, t_i 为 y 变量的相同秩次数。然后就是对 r_s 进行显著性检验。

3. Kendall 相关系数

Kendall 等级相关系数 (Kendall tau rank correlation coefficient) 与 Spearman 相关系数一样, 也是利用“等级”来研究两个变量之间的相关程度, 但考虑的角度却不同。

首先将 n 对配对数据 (x_i, y_i) 评出相应的等级 (R_{xi}, R_{yi}) , 再分别考察 R_{xi} 与 R_{yi} 的一致性 (Concordance)。如果两个等级由小到大排列, 称为一个一致对, 记作 +1, 将 U_x 表示为 R_{xi} 中的一致对的数目, U_y 表示为 R_{yi} 中的一致对的数目; 如果两个等级由大到小排列, 称为一个非一致对, 记作 -1, 将 V_x 表示为 R_{xi} 中的非一致对的数目, V_y 表示为 R_{yi} 中的非一致对的数目。一般会将 R_{xi} 按照自然顺序由小到大排列, 这样 R_{xi} 中的两个等级之间都是一致对。如有这

样的序列： R_{xi} 为 1 2 3 4； R_{yi} 为 4 3 1 2。在 R_{xi} 中等级对 (1,2)(1,3)(1,4)(2,3)(2,4)(3,4) 都是一致的，故 $U_x=6$ ， $V_x=0$ ；在 R_{yi} 中等级对 (4,3)(4,1)(4,2)(3,1)(3,2) 都是非一致对，只有 (1,2) 是一致对，故 $U_y=1$ ， $V_y=5$ 。

在 R_{xi} 按自然顺序排列时， R_{yi} 的一致对最大数目产生于 R_{yi} 也按自然顺序排列，此时它等于 C_n^2 ，用 R_{yi} 的实际一致对数目与最大可能一致对数目相比较，可以测定 x 与 y 的相关程度。

R_{yi} 一致对数目与最大可能一致对数目比表示为

$$\frac{U_y}{C_n^2} = \frac{2U_y}{n(n-1)}$$

R_{yi} 非一致对数目与最大可能非一致对数目之比表示为

$$\frac{V_y}{C_n^2} = \frac{2V_y}{n(n-1)}$$

当 R_{yi} 完全按自然顺序排列时， $\frac{U_y}{C_n^2}$ 的值为 1， $\frac{V_y}{C_n^2}$ 的值为 0；而当 R_{yi} 完全与 R_{xi} 相反时， $\frac{U_y}{C_n^2}$ 的值为 0， $\frac{V_y}{C_n^2}$ 的值为 1。为测定两组等级之间的相关程度，定义的相关系数取值范围为 -1~+1。因此，Kendall 等级相关系数 τ 的定义公式为

$$\tau = \frac{4U_y}{n(n-1)} - 1$$

或

$$\tau = 1 - \frac{4V_y}{n(n-1)}$$

如果 x 与 y 有完全相同的等级，则 $\tau = +1$ ，表明 x 与 y 完全正相关；如果 x 与 y 有完全相反的等级，则 $\tau = -1$ ，表明 x 与 y 完全负相关。一般认为 $|\tau| > 0.8$ ，两组等级相关的程度较高。

11.2.2 双变量过程的参数设置

选择菜单“分析 (Analyze) 相关 (Correlate) 双变量 (Bivariate)”，则弹出如图 11-2 所示对话框 (Bootstrap 过程在这里不再介绍)。

1. 变量设置

进行分析之前要进行变量设置操作，图 11-2 中的左侧是变量列表框，变量 (Variables) 选项框用于选入需要分析的变量，至少要选入两个变量，如果选入变量个数大于两个变量时，则对其分别进行两两分析。

2. 相关系数 (Correlation Coefficients) 栏

此栏用于选择计算的相关系数。

- Pearson：计算 Pearson 相关系数。
- Kendall 的 tau-b：对于分类或者等级变量计算 Kendall 相关系数。
- Spearman：对于分类或者等级变量计算 Spearman 相关系数。

3. 显著性检验 (Test of Significance) 栏

此栏用于定义相关系数的检验方法。

- 双侧检验：(Two-tailed)
- 单侧检验：(One-tailed)

4. 标记显著相关性 (Flag significant correlations)

用 “*” 标记有统计学意义的相关系数，如果 $\text{Sig.} < 0.05$ 则利用一个 “*” 来标记；如果 $\text{Sig.} < 0.01$ 则用两个 “**” 标记。

5. 选项 (Options) 设置

单击 “选项 (Options)” 按钮，则弹出如图 11-3 所示对话框。此对话框用于选择输出统计量和定义缺失值的处理方式。

统计 (Statistics) 栏：选择输出统计量。

- 均值和标准差 (Means and Standard Deviations)：输出各个变量的样本均值及标准差。
- 叉积偏差和协方差 (Cross-product Deviations and Covariances)：输出各对变量的交叉积及协方差矩阵。

缺失值 (Missing Values) 栏：用于定义缺失值。

- 成对排除个案 (Exclude Cases Pairwise)：仅当数据要分析的变量值缺失时才剔除该数据。
- 成列排除个案 (Exclude Cases Listwise)：只要数据中有变量值缺失就剔除该数据。



图 11-2 “双变量 (Bivariate) 设置”对话框



图 11-3 “选项 (Options) 设置”对话框

11.2.3 实例分析



结果文件

——附带光盘“PROGRAM\CH11\实例 11-1”文件夹



动画演示

——附带光盘“AVI\实例 11-1.avi”文件

本实例利用 SPSS 中自带的数据集 car_sales.sav 来进行两变量的相关分析，此数据集是关于汽车销售方面的数据集，数据集共包括 model、sales、resale、type、price 等 26 个变量。数据集 car_sales.sav 的格式如图 11-4 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|----------|-----|----|------|---------------------|----------------|----|----|----|----|----|
| 1 | manufact | 字符串 | 13 | 0 | Manufacturer | 无 | 无 | 13 | 左 | 名义 | 输入 |
| 2 | model | 字符串 | 17 | 0 | Model | 无 | 无 | 17 | 左 | 名义 | 输入 |
| 3 | sales | 数字 | 11 | 3 | Sales in thousa... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 4 | resale | 数字 | 11 | 3 | 4-year resale va... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 5 | type | 数字 | 11 | 0 | Vehicle type | {0, Automob... | 无 | 8 | 右 | 有序 | 输入 |
| 6 | price | 数字 | 11 | 3 | Price in thousa... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 7 | engine_s | 数字 | 11 | 1 | Engine size | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 8 | horsepow | 数字 | 11 | 0 | Horsepower | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 9 | wheelbas | 数字 | 11 | 1 | Wheelbase | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 11-4 数据集 car_sales.sav 的格式

1. 参数设置

选择菜单“分析 (Analyze) 相关 (Correlate) 双变量 (Bivariate)”，则弹出如图 11-5 所示对话框，选中变量 Sales in thousands 和 Fuel efficiency 进入“变量 (Variables)”选项栏中，然后单击“确定”按钮进行相关分析。



图 11-5 “双变量 (Bivariate) 参数设置”对话框

2. 结果分析

设置好参数以后,则单击“确定”按钮进行系统的相关分析,结果如图 11-6 所示,图中给出了 Pearson 相关系数。从检验结果可以看出 Sig 值为 0.837,远远大于 0.10,所以,这两个变量之间并无很强的相关性,故设计不应该把注意力放在汽车的油耗上,因为它对汽车的销售量并无显著效果。

| 相关性 | | | |
|--------------------|---------|--------------------|-----------------|
| | | Sales in thousands | Fuel efficiency |
| Sales in thousands | 皮尔逊相关性 | 1 | -.017 |
| | 显著性(双尾) | | .837 |
| | 个案数 | 157 | 154 |
| Fuel efficiency | 皮尔逊相关性 | -.017 | 1 |
| | 显著性(双尾) | .837 | |
| | 个案数 | 154 | 154 |

图 11-6 Pearson 相关系数

也可以通过绘制散点图来观测相关性,选择菜单“图形(Graphs) 图形构建器(Chart Builder)”,则弹出如图 11-7 所示的“图形绘制”对话框。选中“选择范围(Choose form)”选项栏中的“散点图/点图(Scatter/Dot)”选项并选中简单散点图(Simple Scatter),然后选中变量 Sales in thousands 到 Y 轴,选中变量 Fuel efficiency 到 X 轴,结果如图 11-7 所示。

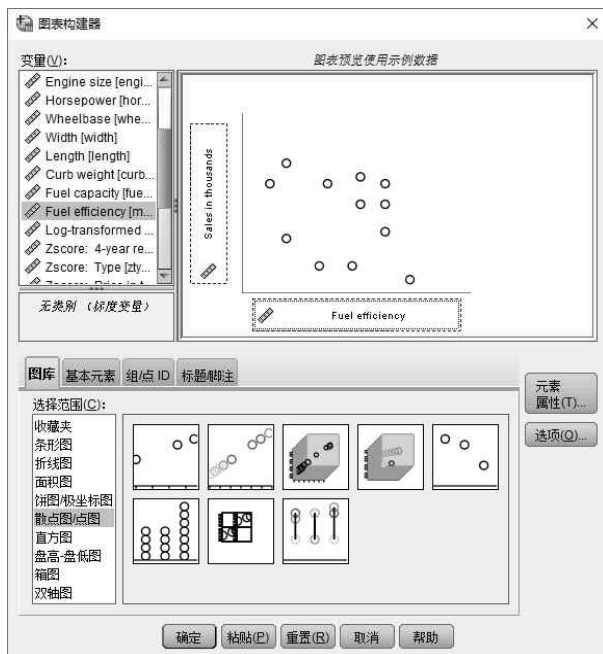


图 11-7 “图形绘制(Plot)”对话框

设置好变量之后单击图 11-7 中的“确定”按钮,则输出图形结果,如图 11-8 所示。从图中的结果可以看出变量 Sales in thousands 的具体分布。

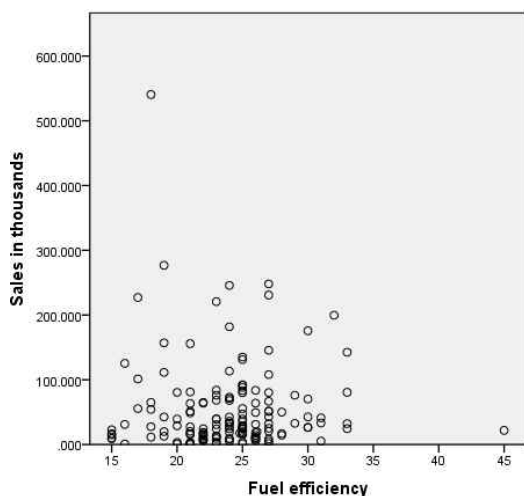


图 11-8 图形绘制结果

11.3 偏相关 (Partial) 过程

偏相关分析即当有多个变量存在时,为了研究任何两个变量之间的关系,而使与这两个变量有联系的其他变量都保持不变。即控制了其他一个或多个变量的影响下,计算两个变量的相关性。

偏相关系数用来衡量任何两个变量之间关系的大小。

11.3.1 偏相关过程的参数设置

选择菜单“分析 (Analyze) 相关 (Correlate) 偏相关 (Partial)”,弹出如图 11-9 所示对话框,此对话框用来设置偏相关分析相关参数。

1. 变量选择设置

图 11-9 中左边为变量列表,变量框 (Variables) 用于选择要进行偏相关分析的变量,至少选入两个变量,如果选入的变量个数大于两个,则系统会分别进行两两相关分析。控制 (Controlling for) 变量框用于选择偏相关分析中的控制变量,如果不选的话,则等同于进行一般的相关分析。

2. 显著性检验 (Test of Significance) 栏

此栏用于定义相关系数的检验方法。

- 双尾检验 (Two-tailed)
- 单尾检验 (One-tailed)

3. 显示实际显著性水平 (Display Actual Significance Level) 栏

选择是否给出真实的显著性水平值,系统默认选项。

4. 选项 (Options) 设置

单击图 11-9 中的“选项 (Options)”按钮, 则弹出如图 11-10 所示对话框, 此对话框用于选择输出统计量和定义缺失值的处理方式。

统计量 (Statistics) 栏: 选择输出统计量。

- 均值和标准差 (Means and standard deviations): 输出各个变量的样本均值及标准差。
- 零阶相关性 (Zero-order correlations): 输出控制变量在内的所有变量的相关矩阵。

缺失值 (Missing Values) 栏: 用于定义缺失值。

- 成列排除个案 (Exclude cases pairwise): 仅当数据要分析的变量值缺失时才剔除该数据。
- 成对排除个案 (Exclude cases listwise): 只要数据中有变量值缺失就剔除该数据。



图 11-9 “偏相关 (Partial) 设置”对话框



图 11-10 “选项 (Options) 设置”对话框

11.3.2 实例分析



结果文件——附带光盘“PROGRAM\CH11\实例 11-2”文件夹



动画演示——附带光盘“AVI\实例 11-2.avi”文件

本实例利用 SPSS 中自带的数据集 health_funding.sav 进行偏相关分析, 包含 funding、disease、visits, 以及 citycode 变量, 数据集 health_funding.sav 的数据格式如图 11-11 所示。下面就分析这些变量之间的偏相关关系。

| health_funding.sav [数据集1] - IBM SPSS Statistics 数据编辑器 | | | | | | | | | | | | |
|---|----------|----|----|------|----------------------|---|----|---|----|----|----|--|
| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 | |
| 1 | funding | 数字 | 8 | 2 | Health care fun... | 无 | 无 | 8 | 右 | 标度 | 输入 | |
| 2 | disease | 数字 | 8 | 2 | Reported disea... | 无 | 无 | 8 | 右 | 标度 | 输入 | |
| 3 | visits | 数字 | 8 | 2 | Visits to health ... | 无 | 无 | 8 | 右 | 标度 | 输入 | |
| 4 | citycode | 数字 | 2 | 0 | | 无 | 无 | 8 | 右 | 标度 | 输入 | |

图 11-11 数据集 health_funding.sav 的数据格式

1. 参数设置

选择菜单“分析 (Analyze) 相关 (Correlate) 偏相关 (Partial)”, 弹出如图 11-12 所示对话框, 此对话框用来设置偏相关分析相关参数。选中变量 Health care funding 和 Reported disease rate 并选入到“变量 (Variables)”选项栏中, 选中变量 Visits to health care providers 并选入到“控制 (Controlling for)”选项栏中。

然后单击图 11-12 中的“选项 (Options)”按钮, 弹出如图 11-13 所示对话框, 选中“零阶相关系数 (Zero-order Correlations)”选项栏和“按列表排除个案 (Exclude Cases Listwise)”选项栏, 接着单击“继续 (Continue)”按钮返回主界面。

返回到偏相关 (Partial Correlations Dialog) 主界面, 单击“确定”按钮运行偏相关分析过程。

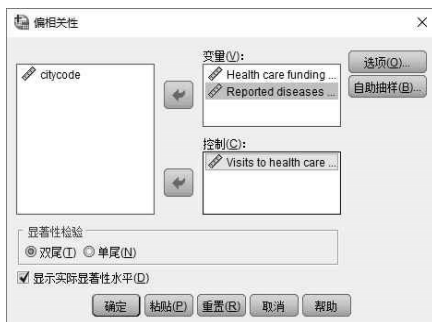


图 11-12 “偏相关 (Partial) 设置”对话框



图 11-13 “选项 (Options) 设置”对话框

2. 结果分析

运行偏相关分析过程后, 系统输出结果, 如图 11-14 所示为偏相关分析表。

| 相关性 | | | | | |
|---|---|----------|---|--|--|
| 控制变量 | | | Health care funding (amount per 100) | Reported diseases (rate per 10,000) | Visits to health care providers (rate per 10,000) |
| 无 ^a | Health care funding (amount per 100) | 相关性 | 1.000 | .737 | .964 |
| | | 显著性 (双尾) | . | .000 | .000 |
| | | 自由度 | 0 | 48 | 48 |
| | Reported diseases (rate per 10,000) | 相关性 | .737 | 1.000 | .762 |
| | | 显著性 (双尾) | .000 | . | .000 |
| | | 自由度 | 48 | 0 | 48 |
| Visits to health care providers (rate per 10,000) | Health care funding (amount per 100) | 相关性 | .964 | .762 | 1.000 |
| | | 显著性 (双尾) | .000 | .000 | . |
| | | 自由度 | 48 | 48 | 0 |
| | Reported diseases (rate per 10,000) | 相关性 | .013 | .928 | |
| | | 显著性 (双尾) | . | .000 | |
| | | 自由度 | 0 | 47 | |

a. 单元格包含零阶 (皮尔逊) 相关性。

图 11-14 偏相关分析表

从图 11-14 输出结果可以看到，变量 Health care funding 和 Reported disease rates 的零阶相关系数为 0.737，其检验结果为 0.000 远远小于 0.001。但是变量 Health care funding 和 disease rates 的一阶相关系数为 0.013，其 P 值为 0.928 远远大于 0.001，所以，其相关关系并不显著。所以，这两个变量之间不可以简单的判断是否具有相关关系。

11.4 Distances (距离) 过程

距离是对观测量之间或者变量之间的相似或不相似程度程度的测度，距离分析可以计算观测之间的距离大小，距离分析不会给出常用的显著性 P 值，只给出各个变量之间的距离大小。当然距离分析的结果也可以应用到聚类分析、因子分析等多元分析方法之中。

距离分析中常用的距离如下。

- 欧几里得距离： $d_{ij}(2) = \left(\sum_{a=1}^p (x_{ia} - x_{ja})^2 \right)^{1/2}$
- 绝对距离： $d_{ij}(1) = \sum_{a=1}^p |x_{ia} - x_{ja}|$
- 切比雪夫距离： $d_{ij}(\infty) = \max_{1 \leq a \leq p} |x_{ia} - x_{ja}|$
- 闵科夫斯基距离： $d_{ij}(q) = \left(\sum_{a=1}^p |x_{ia} - x_{ja}|^q \right)^{1/q}$
- 马尔科夫氏距离

设指标的协方差矩阵 $\Sigma = (\sigma_{ij})_{p \times p}$ ，其中， $\sigma_{ij} = \frac{1}{n-1} \sum_{a=1}^n (x_{ai} - \bar{x}_i)(x_{aj} - \bar{x}_j)$ ，

$\bar{x}_i = \frac{1}{n} \sum_{a=1}^n x_{ai}$ ， $\bar{x}_j = \frac{1}{n} \sum_{a=1}^n x_{aj}$ ， $i, j = 1, \dots, p$ ，如果 Σ^{-1} 存在，则两个样品之间的马氏距离为

$$d_{ij}^2(M) = (X_i - X_j)' \Sigma^{-1} (X_i - X_j)$$

在上述定义的距离中，每种距离的适应情况是不一样的，例如，当各变量的测量值相差悬殊时，要用闵科夫斯基距离并不合理，常需要先对数据标准化，然后用标准化后的数据计算距离。

11.4.1 Distances 过程的距离分析参数设置

选择菜单“分析 (Analyze) 相关 (Correlate) 距离 (Distances)”，弹出如图 11-13 所示对话框，此对话框用来设置距离分析相关参数。

1. 变量选择设置

图 11-15 的左边是变量列表框，变量 (Variables) 选项框用于选择要进行距离分析的变量，至少要选入两个变量。标注个案 (Label Cases by) 选项栏用来选择标识变量。

2. 计算距离 (Compute Distances) 选项栏

用来定义距离分析类型。

- 个案间 (Between Cases) : 定义对观测值进行距离分析。
- 变量间 (Between Variables) : 定义对变量进行距离分析。

3. 测量标准 (Measure) 选项栏

选择距离分析的测度类型。

- 零相似性 (Dissimilarities) : 计算不相似性测度。
- 相似性 (Similarities) : 计算相似性测度。
- 测量 (Measures) 按钮 : 如果要选择不相似性, 单击此按钮, 则弹出如图 11-16 所示的

“测量 (Measures)”对话框, 用于定义距离分析的测度类型。如果要选择相似性, 单击此按钮, 则弹出如图 11-17 所示的“测量 (Measures)”对话框。



图 11-15 “距离分析设置”对话框



图 11-16 “测量 (Measures)”对话框 1



图 11-17 “测量 (Measures)”对话框 2

首先是不相似测度下的“测量 (Measures)”对话框, 如图 11-16 所示。此对话框主要有如下几部分组成。

(1) 测量 (Measure) 选项栏

此栏用于根据变量或观测值数据类型的不同, 选择不同的相似测度即距离测度指标。

- 区间 (Interval) 选项 : 用于计算定距变量的计算方法, 其测量 (Measure) 下拉菜单如图 11-18 所示。包含的计算方法有欧氏距离 (Euclidean distance)、平方欧氏距离 (Squared Euclidean distance)、切比雪夫距离 (Chebychev)、块距离 (Block)、明可夫斯基距离 (Minkowski)、定制距离 (Customized)。
- 计数 (Counts) 选项栏 : 用于设置计数变量, 其下拉菜单如图 11-19 所示。包括两种方法, 卡方测量 (Chi-square measure) 和 Phi 平方测量 (Phi-square measure)。



图 11-18 “区间 (Interval)” 选项



图 11-19 Counts 选项

- 二元 (Binary) 选项栏：首先需要指定表征特征存在与否的取值，在制定测度计算方法。其下拉菜单如图 11-20 所示。其下的 Present 输入框指定表征特征存在的变量值，默认为 1；Absent 输入框指定表征特征不存在的变量值，默认为 0。包含的计算方法有欧氏距离 (Euclidean distance)、平方欧氏距离 (Squared Euclidean distance)、大小差测度 (Size difference)、模式差异测度 (Pattern difference)、方差测度 (Variance)、形状测度 (Sharp)、兰斯 - 威廉姆斯测度 (Lance and Willians)。

(2) 转换值 (Transform Values) 选项栏

此栏用于定义标准化方法，当数据的量纲不一致时会使得分析结果发生偏差，此时有必要对数据进行标准化。其下拉菜单如图 11-21 所示。

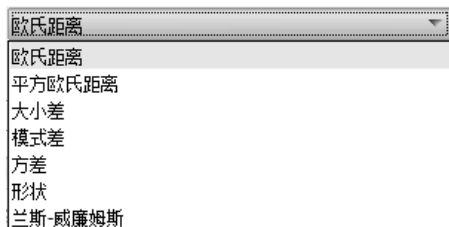


图 11-20 “二元 (Binary) 选项设置” 对话框



图 11-21 标准化方法选择菜单

图 11-21 所示下拉菜单中包含的方法主要有如下几种。

- 无 (None)。
- Z 得分 (Z scores)：进行 Z 变换。
- 范围 -1 到 1 (Range -1 to 1)：将数据范围标准化到 -1 ~ 1。
- 范围 0 到 1 (Range 0 to 1)：将数据范围标准化到 0 ~ 1。
- 最大量级为 1 (Maximum magnitude of 1)：将数据标准化后使得其最大值为 1。
- 平均值为 1 (Mean of 1)：将数据标准化后使得其平均值为 1。
- 标准差为 1 (Standard deviation of 1)：将数据标准化后使得其标准差为 1。

进行完标准化数据以后，还应该选择是对变量进行标准化 (By Variable)，还是对观测值进行标准化 (By Case)。

(3) 转换测量 (Transform Measures) 选项栏

此栏用于定义对计算出来的距离测度做进一步的转化，方法如下。

- 绝对值 (Absolute Values)。
- 变化量符号 (Change Sign)。
- 重新标度到 0 ~ 1 范围 (Rescale to 0-1 Range)。

然后是相似测度下的“测量 (Measures)”对话框，如图 11-17 所示。此对话框和如图 11-16

所示的“测量 (Measures)”对话框大致相似,不同的是在测量 (Measures) 选项栏中的区间 (Interval) 单选框中的设置不同,如图 11-22 所示。

- 皮尔逊相关性 (Pearson): 皮尔逊相关系数。
- 余弦: 角度相似系数。

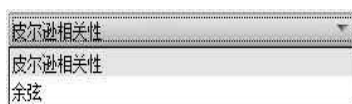


图 11-22 区间下拉菜单

11.4.2 实例分析



结果文件

——附带光盘“PROGRAM\CH11\实例 11-3”文件夹



动画演示

——附带光盘“AVI\实例 11-3.avi”文件

本实例考虑数据集为 CH1103.sav, 此数据集包含全国各个省市自治区直辖市的农民家庭收支的分布规律, 数据集包含 7 个变量分别是地区、食品、衣着、燃料、住房、生活用品以及文化生活, 下面就分析各个指标变量之间的距离相关关系, 如图 11-23 所示是所要分析的数据集的格式。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|------|-----|----|------|----|---|----|----|----|----|----|
| 1 | 地区 | 字符串 | 9 | 0 | | 无 | 无 | 6 | 左 | 名义 | 输入 |
| 2 | 食品 | 数字 | 11 | 1 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 3 | 衣着 | 数字 | 11 | 1 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 4 | 燃料 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 5 | 住房 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 6 | 生活用品 | 数字 | 11 | 1 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 7 | 文化生活 | 数字 | 11 | 2 | | 无 | 无 | 11 | 右 | 标度 | 输入 |

图 11-23 数据集 CH1103.sav 的格式

1. 参数设置

选择菜单“分析 (Analyze) 相关 (Correlate) 距离 (Distances)”, 弹出如图 11-24 所示对话框, 此对话框用来设置距离分析相关参数。选择变量食品、衣着、燃料、住房、生活用品, 以及文化生活到“变量 (Variables)”选项栏中, 在“计算距离 (Compute Distance)”选项栏中选择“变量间 (Between Variables)”选项, “测量标准 (Measure)”选项栏中选择“相似性 (Similarities)”选项。

然后单击“测量 (Measures)”按钮, 弹出如图 11-25 所示对话框, 在“标准化 (Standardize)”下拉菜单中选择 Z 得分 (Z Scores) 变量, 然后单击“继续 (Continue)”按钮返回主界面。



图 11-24 “距离 (Distances) 设置”对话框



图 11-25 “测量 (Measures) 设置”对话框

2. 结果分析

返回主界面后，单击“确定”按钮进行距离分析，结果如图 11-26 所示为基本统计信息。

| 个案处理摘要 | | | | | |
|--------|--------|----------|------|-----|--------|
| 有效 | | 个案 缺失 | | 总计 | |
| 个案数 | 百分比 | 个案数 | 百分比 | 个案数 | 百分比 |
| 28 | 100.0% | 0 | 0.0% | 28 | 100.0% |

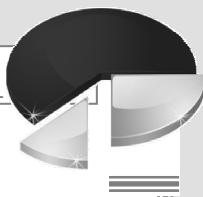
图 11-26 基本统计信息

然后输出的是变量之间的近似矩阵，如图 11-27 所示。近似矩阵是标准的对称矩阵，通过近似矩阵可以观察到变量之间的距离关系的强弱，例如，住房和生活用品之间的距离相关系数为 0.843，说明此两个变量之间的距离关系很强。而变量文化生活和变量衣着之间的距离相关系数为 0.181，为最小，说明此两个变量之间的距离相关强度最弱。

| 近似值矩阵 | | | | | | |
|------------|-------|-------|-------|-------|-------|-------|
| 值的向量之间的相关性 | | | | | | |
| | 食品 | 衣着 | 燃料 | 住房 | 生活用品 | 文化生活 |
| 食品 | 1.000 | .518 | .517 | .778 | .707 | .613 |
| 衣着 | .518 | 1.000 | .133 | .579 | .752 | .181 |
| 燃料 | .517 | .133 | 1.000 | .133 | .210 | .456 |
| 住房 | .778 | .579 | .133 | 1.000 | .843 | .353 |
| 生活用品 | .707 | .752 | .210 | .843 | 1.000 | .336 |
| 文化生活 | .613 | .181 | .456 | .353 | .336 | 1.000 |

这是相似性矩阵

图 11-27 近似矩阵



第 12 章 聚类分析

将物理或抽象对象的集合分组为由类似的对象组成的多个类的过程称为聚类。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。在许多应用中，可以将一个簇中的数据对象作为一个整体来对待。

聚类分析是一种重要的人类行为。人们通过不断地改进下意识中的聚类模式来学会如何区分猫和狗，或者动物和植物。聚类分析已经广泛地应用于模式识别、数据分析、图像处理，以及市场研究等学科中。通过聚类，人们能够识别密集的和稀疏的区域，因而发现全局的分布模式，以及数据属性之间的有趣的相互关系。

聚类分析的广泛应用，在各种社会经济、商业策划、技术研发等活动中随处可见，本章将详细叙述利用 SPSS 软件系统进行聚类分析。



本讲内容

- 聚类分析的原理简介
- 快速样本聚类过程
- 系统聚类过程
- 二阶聚类分析

12.1 聚类分析的原理

聚类分析又称群分析，它是研究（样品或指标）分类问题的一种多元统计方法。类，通俗地说，就是指相似元素的集合。严格的数学定义是较麻烦的，在不同问题中类的定义是不同的。

聚类分析起源于分类学，随着生产技术和科学的发展，人类的认识不断加深，分类越来越细，要求也越来越高，有时光凭经验和专业知识是不能进行确切分类的，往往需要定性和定量分析结合起来去分类，于是数学工具逐渐被引进分类学中，形成了数值分类学。后来随着多元分析的引进，聚类分析又逐渐从数值分类学中分离出来而形成相对独立的分支。

在社会经济领域中存在着大量分类问题，如对我国省市自治区经济发展水平的分析，又如若对某些大城市的物价指数进行考察，而物价指数很多，有农用生产物价指数、服务

项目价指数、食品消费物价指数、建材零售价格指数等。由于要考察的物价指数很多，通常先对这些物价指数进行分类。总之，需要分类的问题很多，因此，聚类分析这个有用的数学工具越来越受到人们的重视，它在许多领域中都得到了广泛的应用。

聚类分析包含系统聚类法、样品聚类法、动态聚类法、模糊聚类法、图论聚类法、聚类预报法等。

12.1.1 一般原理

在进行聚类分析之前，要对聚类分析的基本概念和基本原理有所了解，由上可知，聚类分析又称群分析，其属于无监督的学习方法，是将对象集划分为若干类别的过程。

所以，为了将样品（或指标）进行分类，就需要研究样品之间关系。聚类分析中主要采用如下两种方法。

相似系数法，即性质越接近的样品，它们的相似系数的绝对值越接近 1，而彼此无关的样品，它们的相似系数的绝对值越接近于 0。比较相似的样品归为一类，不怎么相似的样品归为不同的类。

距离法，即将一个样品看作 p 维空间的一个点，并在空间定义距离，距离较近的点归为一类，距离较远的点归为不同的类。

由于实际问题中，遇到的指标有的是定量的（如长度、重量等），有的是定性的（如性别、职业等），因此，将变量（指标）的类型按以下三种尺度划分。

- 间隔尺度：变量是用连续的量来表示的，如长度、重量、压力、速度等。在间隔尺度中，如果存在绝对零点，又称比例尺度。
- 有序尺度：变量度量时没有明确的数量表示，而是划分一些等级，等级之间有次序关系，如某产品分上、中、下三等，此三等有次序关系，但没有数量表示。
- 名义尺度：变量度量时、既没有数量表示，也没有次序关系，如某物体有红、黄、白三种颜色，市场供求中的“产”和“销”等。

不同类型的变量，在定义距离和相似系数时，其方法有很大差异，使用时必须注意。研究比较多的是间隔尺度。另外，要弄清 Q 型（样本）聚类和 R 型（变量）聚类。Q 型聚类即是要把所有观测记录进行分类，把性质相近的观测分在同一类中，性质差异较大的观测分在不同的类的过程。R 型聚类即是要把变量进行分类的过程。下面主要对 Q 型聚类分析进行介绍。

设有 n 个样品，每个样品测得 p 项指标（变量），原始资料阵为

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

其中， $x_{ij} (i=1, \cdots, n, j=1, \cdots, p)$ 为第 i 个样品的第 j 个指标的观测数据。第 i 个样品 X_i 为矩阵 X 的第 i 行描述，所以，任何两个样品 X_K 与 X_L 之间的相似性，可以通过矩阵 X 中的第 K 行与第 L 行的相似程度来刻画；任何两个变量 x_K 与 x_L 之间的相似性，可以通过第 K 列与

第 L 列的相似程度来刻画。

1. 距离

由距离法可知, 聚类分析之前首先要定义距离, 设 n 个样品 (X 中的 n 个行) 看作 p 维空间中 n 个点, 则两个样品间相似程度可用 p 维空间中两点的距离来度量。令 d_{ij} 表示样品 X_i 与 X_j 的距离, 聚类分析中的常用距离有下述几种。

- 欧几里得距离: $d_{ij}(2) = \left(\sum_{a=1}^p (x_{ia} - x_{ja})^2 \right)^{1/2}$
- 绝对距离: $d_{ij}(1) = \sum_{a=1}^p |x_{ia} - x_{ja}|$
- 切比雪夫距离: $d_{ij}(\infty) = \max_{1 \leq a \leq p} |x_{ia} - x_{ja}|$
- 闵科夫斯基距离: $d_{ij}(q) = \left(\sum_{a=1}^p |x_{ia} - x_{ja}|^q \right)^{1/q}$
- 马尔科夫距离: 设 Σ 表示指标的协差阵 $\Sigma = (\sigma_{ij})_{p \times p}$, 其中 $\sigma_{ij} = \frac{1}{n-1} \sum_{a=1}^n (x_{ai} - \bar{x}_i)(x_{aj} - \bar{x}_j)$, $\bar{x}_i = \frac{1}{n} \sum_{a=1}^n x_{ai}$, $\bar{x}_j = \frac{1}{n} \sum_{a=1}^n x_{aj}$ ($i, j = 1, \dots, p$), 如果 Σ^{-1} 存在, 则两个样品之间的马马尔科夫距离为

$$d_{ij}^2(M) = (X_i - X_j)' \Sigma^{-1} (X_i - X_j)$$

这里 X_i 为样品 X_i 的 p 个指标组成的矢量, 即原始资料阵的第 i 行矢量。样品 X_j 类似。

顺便给出样品 X 到总体 G 的马氏距离定义为

$$d^2(X, G) = (X - \mu)' \Sigma^{-1} (X - \mu)$$

式中, μ 为总体的均值矢量; Σ 为协方差阵。

- 兰氏距离: 它是由 Lance 和 Williams 最早提出的, 故称兰氏距离, 公式如下:

$$d_{ij}(L) = \frac{1}{p} \sum_{a=1}^p \frac{|x_{ia} - x_{ja}|}{x_{ia} + x_{ja}}, \quad i, j = 1, \dots, n$$

在上述定义的距离中, 每种距离的适应情况是不一样的, 例如, 当各变量的测量值相差悬殊时, 要用闵科夫斯基距离并不合理, 常需要先对数据标准化, 然后用标准化后的数据计算距离。闵科夫斯基距离不足之处主要有以下两点。

- 距离与各指标的量纲有关。
- 没有考虑指标之间的相关性。

其实, 绝对距离、欧几里得距离和切比雪夫距离都是闵科夫斯基距离的特殊形式, 所以其同样有上述不足之处。

马尔科夫距离的特点是既排除了各指标之间相关性的干扰, 而且还不受各指标量纲的影响。除此之外, 它还有一些优点, 如可以证明, 将原数据进行线性交换后, 马尔科夫距离仍不变等。

兰氏距离的不足是仅适用于一切 $x_{ij} > 0$ 的情况, 这个距离有助于克服各指标之间量纲

的影响,但没有考虑指标之间的相关性。

这里要特别注意的是,以上所定义的距离是适用于间隔尺度变量的,当变量是有序尺度或名义尺度变量时,可以定义其他形式的距离。

2. 相似距离

研究样品之间的关系,除了利用上述距离之外,还有相似系数,相似系数是描写样品之间相似程度的一个量,聚类分析中常用的相似系数有如下几种。

(1) 夹角余弦

将任何两个样品 X_i 与 X_j 看作 p 维空间的两个矢量,这两个矢量的夹角余弦用 $\cos \theta_{ij}$ 表示。则

$$\cos \theta_{ij} = \frac{\sum_{a=1}^p x_{ia} x_{ja}}{\sqrt{\sum_{a=1}^p x_{ia}^2 \cdot \sum_{a=1}^p x_{ja}^2}}, \quad -1 \leq \cos \theta_{ij} \leq 1$$

$\cos \theta_{ij} = 1$, 说明两个样品 X_i 与 X_j 完全相似; $\cos \theta_{ij}$ 接近 1, 说明 X_i 与 X_j 相似密切; $\cos \theta_{ij} = 0$, 说明 X_i 与 X_j 完全不一样; $\cos \theta_{ij}$ 接近 0, 说明 X_i 与 X_j 差别大。由上面的计算可以得到相似系数矩阵如下:

$$\begin{bmatrix} \cos \theta_{11} & \cos \theta_{12} & \cdots & \cos \theta_{1n} \\ \cos \theta_{21} & \cos \theta_{22} & \cdots & \cos \theta_{2n} \\ \vdots & \vdots & & \vdots \\ \cos \theta_{n1} & \cos \theta_{n2} & \cdots & \cos \theta_{nn} \end{bmatrix}$$

式中, $\cos \theta_{11} = \cos \theta_{22} = \cdots = \cos \theta_{nn} = 1$ 。所以,只需计算上三角形部分或下三角形部分,再根据如上相似矩阵可对 n 个样品进行分类,把比较相似的样品归为一类,否则归为不同的类。

(2) 相关系数

一般指变量间的相关系数,作为刻画样品间的相似关系也可类似给出定义,即第 i 个样品与第 j 个样品之间的相关系数定义为

$$r_{ij} = \frac{\sum_{a=1}^p (x_{ia} - \bar{x}_i)(x_{ja} - \bar{x}_j)}{\sqrt{\sum_{a=1}^p (x_{ia} - \bar{x}_i)^2 \cdot \sum_{a=1}^p (x_{ja} - \bar{x}_j)^2}}, \quad -1 \leq r_{ij} \leq 1$$

式中, $\bar{x}_i = \frac{1}{p} \sum_{a=1}^p x_{ia}$; $\bar{x}_j = \frac{1}{p} \sum_{a=1}^p x_{ja}$ 。

实际上, r_{ij} 就是两个矢量 $X_i - \bar{X}_i$ 与 $X_j - \bar{X}_j$ 的夹角余弦,其中

$$\bar{X}_i = (\bar{x}_i, \cdots, \bar{x}_i)', \quad \bar{X}_j = (\bar{x}_j, \cdots, \bar{x}_j)'$$

若将原始数据标准化,则 $\bar{X}_i = \bar{X}_j = 0$, 这时 $r_{ij} = \cos \theta_{ij}$ 。

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}$$

式中, $r_{11} = r_{22} = \cdots = r_{nn} = 1$, 可根据 R 对 n 个样品进行分类。

12.1.2 聚类分析步骤

了解了聚类分析的基本原理之后来叙述一下聚类分析主要包含的四个步骤, 第一要根据研究目标确定合适的聚类变量; 第二要计算距离和相似度测度; 第三选定聚类方法进行聚类分析; 第四对聚类结果进行分析。

1. 数据预处理

数据预处理包括选择数量、类型和特征的标度, 它依靠特征选择和特征抽取, 特征选择选择重要的特征, 特征抽取把输入的特征转化为一个新的显著特征, 它们经常被用来获取一个合适的特征集来避免“维数灾”进行聚类, 数据预处理还包括将孤立点移出数据, 孤立点是不依附于一般数据行为或模型的数据, 因此, 孤立点经常会导致有偏差的聚类结果, 因此, 为了得到正确的聚类, 必须将它们剔除。

由聚类分析基本原理知其基础是根据所选变量对研究对象进行聚类分析, 聚类分析结果仅仅反映了所选变量所定义的数据结构, 所以, 变量选择在聚类分析中非常重要。选择变量应该根据所研究对象的特征来选择, 总之, 所选择的便利应该具有如下一些特点。

- 和聚类目标相关性高。
- 反映分类对象特征。
- 不同的研究对象上具有较大差异。
- 变量之间相关性较低。

2. 相似度计算

根据聚类对象, 选择相应的相似度方法和距离定义公式进行计算, 计算样品或者变量之间的距离, 以及类与类之间的距离。

3. 聚类或分组

聚类过程涉及两个重要问题。

选择聚类方法, 不同的聚类方法, 得到的聚类结果往往是不同的, 最常见的聚类方法有系统聚类法、快速聚类法、分层聚类法和二阶聚类法。

分类数的确定, 如何确定分类数的多少在聚类分析中也很重要, 往往需要考虑实际案例中的分类要求和特点等。

将数据对象分到不同的类中是一个很重要的步骤, 数据基于不同的方法被分到不同的类中, 划分方法和层次方法是聚类分析的两个主要方法, 划分方法一般从初始划分和最优化一个聚类标准开始。Crisp Clustering, 它的每一个数据都属于单独的类; Fuzzy

Clustering, 它的每个数据可能在任何一个类。Crisp Clustering 和 Fuzzy Clustering 是划分方法的两个主要技术, 划分方法聚类是基于某个标准产生一个嵌套的划分系列, 它可以度量不同类之间的相似性或一个类的可分离性用来合并和分裂类, 其他的聚类方法还包括基于密度的聚类、基于模型的聚类、基于网格的聚类。

4. 结果分析

评估聚类结果是另一个重要的阶段, 聚类是一个无管理的程序, 也没有客观的标准来评价聚类结果, 它通过一个类有效索引来评价, 类有效索引在决定类的数目时经常扮演一个重要角色, 类有效索引的最佳值被期望从真实的类数目中获取, 一个通常的决定类数目的方法是选择一个特定的类有效索引的最佳值, 这个索引能否真实地得出类的数目是判断该索引是否有效的标准, 很多已经存在的标准对于相互分离的类数据集都能得出很好的结果, 但是对于复杂的数据集, 却通常行不通, 例如, 对于交迭类的集合。

12.1.3 系统聚类方法

类与类之间不同的定义距离, 就产生了不同的系统聚类方法。系统聚类方法主要有最短距离法、最长距离法、中间距离法、重心法、类平均法、可变类平均法、可变法、离差平方和法。系统聚类分析方法的步骤基本上是一致的。设 d_{ij} 表示样品 X_i 与 X_j 之间的距离, D_{ij} 表示类 G_i 与 G_j 之间的距离。则系统聚类方法如下。

1. 最短距离法

定义类 G_i 与 G_j 之间的距离为两类最近样品的距离, 即

$$D_{ij} = \min_{G_i \in G_i, G_j \in G_j} d_{ij}$$

设类 G_p 与 G_q 合并成一个新类记为 G_r , 则任一类 G_k 与 G_r 的距离为

$$\begin{aligned} D_{kr} &= \min_{X_i \in G_i, X_j \in G_j} d_{ij} \\ &= \min \left\{ \min_{X_i \in G_k, X_j \in G_p} d_{ij}, \min_{X_i \in G_k, X_j \in G_q} d_{ij} \right\} \\ &= \min \{ D_{kp}, D_{kq} \} \end{aligned}$$

最短距离法聚类步骤如下。

(1) 定义样品之间距离, 计算样品距离, 得到距离阵为 $D_{(0)}$, 开始每个样品自成一类, 显然这时 $D_{ij} = d_{ij}$ 。

(2) 找出 $D_{(0)}$ 的非对角线最小元素, 设为 D_{pq} , 则将 G_p 和 G_q 合并成一个新类, 记为 G_r , 即 $G_r = \{G_p, G_q\}$ 。

(3) 给出计算新类与其他类的距离公式: $D_{kr} = \min \{D_{kp}, D_{kq}\}$, 将 $D_{(0)}$ 中第 p 、 q 行及 p 、 q 列利用上面公式合并成一个新行新列, 新行新列对应 G_r , 所得到的矩阵记为 $D_{(1)}$ 。

(4) 对 $D_{(1)}$ 重复上述对 $D_{(0)}$ 的 (2) (3) 两步得 $D_{(2)}$; 如此下去, 直到所有的元素合并

成一类为止。

注意：如果某一步 $D_{(k)}$ 中非对角线最小的元素不止一个，则对应这些最小元素的类可以同时合并。

2. 最长距离法

定义类 G_i 与类 G_j 之间距离为两类最远样品的距离，即

$$D_{pq} = \max_{X_i \in G_p, X_j \in G_q} d_{ij}$$

最长距离法与最短距离法的并类步骤完全一样，也是将各样品先自成一类，然后将非对角线上最小元素对应的两类合并。设某一步将类 G_p 与 G_q 合并为 G_r ，则任一类 G_k 与 G_r 的距离用最长距离公式为

$$\begin{aligned} D_{kr} &= \max_{X_i \in G_k, X_j \in G_r} d_{ij} \\ &= \max \left\{ \max_{X_i \in G_k, X_j \in G_p} d_{ij}, \max_{X_i \in G_k, X_j \in G_q} d_{ij} \right\} \\ &= \max \{ D_{kp}, D_{kq} \} \end{aligned}$$

再找非对角线最小元素的两类并类，直至所有的样品全归为一类为止。

易见最长距离法与最短距离法只有两点不同：一是类与类之间的距离定义不同；二是计算新类与其他类的距离所用的公式不同。

3. 重心法

定义类与类之间距离时，为了体现出每类包含的样品个数给出重心法。重心法定义两类之间的距离就是两类重心之间的距离。设 G_p 和 G_q 的重心（该类样品的均值）分别是 \bar{X}_p 和 \bar{X}_q （注意一般它们是 p 维矢量），则 G_p 和 G_q 之间的距离是 $D_{pq} = d_{\bar{X}_p \bar{X}_q}$ 。

设聚类到某一步， G_p 和 G_q 分别有样品 n_p, n_q 个，将 G_p 和 G_q 合并为 G_r ，则 G_r 内样品个数为 $n_r = n_p + n_q$ ，它的重心是 $\bar{X}_r = \frac{1}{n_r}(n_p \bar{X}_p + n_q \bar{X}_q)$ ，某一类 G_k 的重心是 \bar{X}_k ，它与新类 G_r 的距离（如果最初样品之间的距离采用欧几里得距离）为

$$\begin{aligned} D_{kr}^2 &= d_{\bar{X}_k \bar{X}_r}^2 = (\bar{X}_k - \bar{X}_r)'(\bar{X}_k - \bar{X}_r) \\ &= \left[\bar{X}_k - \frac{1}{n_r}(n_p \bar{X}_p + n_q \bar{X}_q) \right]' \left[\bar{X}_k - \frac{1}{n_r}(n_p \bar{X}_p + n_q \bar{X}_q) \right] \\ &= \bar{X}_k' \bar{X}_k - 2 \frac{n_p}{n_r} \bar{X}_k' \bar{X}_p - 2 \frac{n_q}{n_r} \bar{X}_k' \bar{X}_q \\ &\quad + \frac{1}{n_r^2} (n_p^2 \bar{X}_p' \bar{X}_p + 2 n_p n_q \bar{X}_p' \bar{X}_q + n_q^2 \bar{X}_q' \bar{X}_q) \end{aligned}$$

利用 $\bar{X}_k' \bar{X}_k = \frac{1}{n_r} \left(n_p \bar{X}_k' \bar{X}_k + n_q \bar{X}_k' \bar{X}_k \right)$ 代入上式得

$$\begin{aligned}
 D_{kr}^2 &= \frac{n_p}{n_r} \left(\bar{X}'_k \bar{X}_k - 2\bar{X}'_p \bar{X}_q + \bar{X}'_p \bar{X}_q \right) + \frac{n_q}{n_r} \left(\bar{X}'_k \bar{X}_k - 2\bar{X}'_k \bar{X}_q + \bar{X}'_q \bar{X}_q \right) \\
 &\quad - \frac{n_p n_q}{n_r^2} (\bar{X}'_p \bar{X}_p - 2\bar{X}'_p \bar{X}_q + \bar{X}'_q \bar{X}_q) \\
 &= \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2 - \frac{n_p}{n_r} \frac{n_q}{n_r} D_{pq}^2
 \end{aligned}$$

如上定义当 $n_p = n_q$ 时即为中间距离法的公式。重心法的归类步骤与以上方法基本上一致，所不同的是每合并一次类，就要重新计算新类的重心及各类与新类的距离。

4. 类平均法

重心法虽有很好的代表性，但并未充分利用各样品的信息，因此给出类平均法，它定义两类之间的距离平方为这两类元素两两之间距离平方的平均，即

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{X_i \in G_p} \sum_{X_j \in G_q} d_{ij}^2$$

设聚类到某一步将 G_p 和 G_q 合并为 G_r ，则任一类 G_k 与 G_r 的距离为

$$D_{kr}^2 = \frac{1}{n_k n_r} \sum_{X_i \in G_k} \sum_{X_j \in G_r} d_{ij}^2 = \frac{1}{n_k n_r} \left(\sum_{X_i \in G_k} \sum_{X_j \in G_p} d_{ij}^2 + \sum_{X_i \in G_k} \sum_{X_j \in G_q} d_{ij}^2 \right) = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2$$

类平均法的聚类步骤与上述方法完全类似。

5. 可变类平均法

由于类平均法公式中没有反映 G_p 与 G_q 之间距离 D_{pq} 的影响，所以，给出可变类平均法，此法定义两类之间的距离同上，只是将任一类 G_k 与新类 G_r 的距离改为如下形式：

$$D_{kr}^2 = \frac{n_p}{n_r} (1 - \beta) D_{kp}^2 + \frac{n_q}{n_r} (1 - \beta) D_{kq}^2 + \beta D_{pq}^2$$

式中， β 是可变的且 $\beta > 1$ 。

6. 可变法

此法定义两类之间的距离仍同上，而新类 G_r 与任一类的 G_k 的距离公式为

$$D_{kr}^2 = \frac{1 - \beta}{2} (D_{kp}^2 + D_{kq}^2) + \beta D_{pq}^2$$

式中， β 是可变的，且 $\beta > 1$ 。显然在可变类平均法中取 $\frac{n_p}{n_r} = \frac{n_q}{n_r} = \frac{1}{2}$ ，即为上式。

可变类平均法与可变法的分类效果与 β 的选择关系极大， β 如果接近 1，一般分类效果不好，在实际应用中 β 常取负值。

7. 离差平方和法

假设将 n 个样品分成 k 类： G_1, G_2, \dots, G_k ，用 $X_i^{(t)}$ 表示 G_t 中的第 i 个样品（注意 $X_i^{(t)}$ 是

p 维矢量), n_t 表示 G_t 中的样品个数, $\bar{X}^{(t)}$ 是 G_t 的重心, 则 G_t 中样品的离差平方和为

$$S_t = \sum_{i=1}^{n_t} (X_i^{(t)} - \bar{X}^{(t)})'(X_i^{(t)} - \bar{X}^{(t)})$$

k 个类的类内离差平方和为

$$S = \sum_{t=1}^k S_t = \sum_{t=1}^k \sum_{i=1}^{n_t} (X_i^{(t)} - \bar{X}^{(t)})'(X_i^{(t)} - \bar{X}^{(t)})$$

其基本思想来自于方差分析, 如果分类正确, 同类样品的离差平方和应当较小, 类与类的离差平方和应当较大。具体做法是先将 n 个样品各自成一类, 然后每次缩小一类, 每缩小一类离差平方和就要增大, 选择使 S 增加最小的两类合并直到所有的样品归为一类为止。

12.2 快速样本聚类过程

12.2.1 快速聚类简介

快速样本聚类的方法(K 中心聚类)就是将聚类仅仅进行到指定的类数就停止。进行快速样本聚类分析应当确定最终聚类数, 使聚类发生到该指定类数后停止。为了使聚类过程快速有效, 还可以指定聚类中心点位置, 这样将使聚类过程的迭代次数减少很多。快速聚类过程始终遵照所有样本空间的点与这几个类中心的距离取最小值原则, 进行反复的迭代计算, 最终将各个个案分配到各个类中心所在的类, 迭代计算将停止。另外, 系统还提供了一种更简单的方法, 即用户指定了初始类中心后, 系统只负责分类, 而不再更改这些初始类中心的位置, 最终将各个个案点归类到各个初始类中心。

快速样本聚类的优点是占内存少、计算量小、处理速度快, 特别适合大样本的聚类分析。但是其缺点是应用范围有限, 要求用户制定分类数目(要告知), 只能对观测量(样本)聚类, 而不能对变量聚类, 且所使用的聚类变量必须都是连续性变量。

12.2.2 SPSS 快速聚类的设置

下面将详细叙述 SPSS 中快速聚类分析的设置用法等, 首先打开数据集, 选择菜单“分析(Analyze) 分类(Classify) K 均值聚类(K-Means Cluster)”, 则系统弹出“K 中心聚类法”对话框, 如图 12-1 所示, 下面介绍 K 中心聚类法各个选项的含义。

1. 主面板框的设置

主面板框如图 12-1 所示, 其中的各个设置的含义如下。

- 变量(Variables): 待分析的数值型变量。
- 个案标记依据(Label Cases by): 样品的标签变量。
- 聚类数(Number of Clusters): 指定聚类个数, 系统默认为 2。
- 聚类方法(Method): 分为迭代与分类和仅分类两种, 其中后者用于已知分类中心。
- 聚类中心(Cluster Centers): 分为两个选项, 即读取初始聚类中心(Read Initial from)和写入最终聚类中心(Write Final as)复选框。前一个指的是选择初始类的

方法, 包含打开数据集 (Open Dataset) 和外部数据文件 (External Dataset); 后一个指的是如何保存聚类结果的类中心, 包含新数据集 (New Dataset) 和数据文件 (Data File)。

- 迭代 (Iterate) 设置: 迭代设置框。
- 选项 (Options) 设置。
- 保存 (Save): 结果保存设置框。

2. 迭代 (Iterate) 设置

由图 12-1 的主面板框, 单击“迭代 (Iterate)”按钮, 则系统弹出如图 12-2 所示的“迭代参数设置”对话框, 其中各选项含义如下。

- 最大迭代次数 (Maximum): 最大迭代次数, 系统默认为 10。
- 收敛性标准 (Convergence): 指定 K-Means 算法的收敛依据, 取值范围为 0 ~ 1, 系统默认为 0。
- 使用运行均值 (Use running means): 要求使用可变类平均数。



图 12-1 “K 中心聚类法”对话框

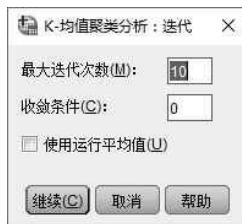


图 12-2 “迭代参数设置”对话框

3. 选项 (Option) 设置

单击图 12-1 中“选项 (Options)”按钮, 则弹出如图 12-3 所示的“选项 (Options) 设置”对话框。

- 统计量 (Statistics): 输出选择的统计量, 框中包含三个选择, 即初始聚类中心 (Initial Cluster Centers)、方差分析表 (ANOVA Table) 和每个个案的聚类信息 (Cluster Information for Each Case)。
- 缺失值 (Missing Values): 缺失值的处理选项, 包括成列排除个案 (Exclude Cases Listwise): 进行多变量分析时, 若某个变量含有缺失值, 则在分析时删

除；成对排除个案 (Exclude Cases Pairwise)：当所有聚类变量都有缺失值时才会将观测删除。

4. 保存 (Save) 设置

在图 12-1 中单击“保存 (Save)”按钮，则弹出如图 12-4 所示的对话框，对话框中有两个选择。

- 聚类成员 (Cluster Membership)：利用一个新变量保存各观测最终被分配到哪一类中，取值为 1 到分类个数。
- 与聚类中心的距离 (Distance from Cluster Center)：利用一个新变量保存各观测最终所属类中心的距离。

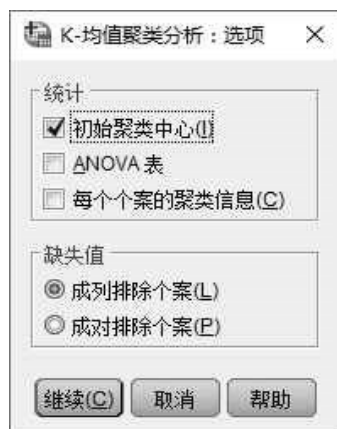


图 12-3 “选项 (Options) 设置”对话框



图 12-4 “保存 (Save) 设置”对话框

12.2.3 实例分析



起始文件

——附带光盘“PROGRAM\Ch12\实例 12-1\数据集 ch1201.xls”



动画演示

——附带光盘“AVI\Ch12\12-1.avi”

1. 数据来源及说明

本案例采用的数据集来源于国家统计局网站，数据集是关于 2006 年全国主要城市空气质量的指标，包括可吸入颗粒物 (mg/m^3)、二氧化硫 (mg/m^3)、二氧化氮 (mg/m^3) 和空气质量达到以及好于二级的天数 (day)，在 SPSS 的操作中分别用 X1、X2、X3 和 X4 来代表上述四种空气质量指标。

2. 分析要求

运用快速聚类分析方法将全国主要城市按不同的尺度分类，以对不同层次的城市空气质量进行分析。分析中的数据参见表 12-1。

表 12-1 2006 年主要城市空气质量指标

| 指 标 | 可吸入颗粒物/(mg/m ³) | 二氧化硫/(mg/m ³) | 二氧化氮/(mg/m ³) | 空气质量达到以及好于二级的天数/day |
|------|-----------------------------|---------------------------|---------------------------|---------------------|
| 北京 | 0.162 | 0.052 | 0.066 | 241 |
| 天津 | 0.114 | 0.067 | 0.048 | 305 |
| 石家庄 | 0.142 | 0.044 | 0.039 | 287 |
| 太原 | 0.142 | 0.080 | 0.025 | 261 |
| 呼和浩特 | 0.102 | 0.054 | 0.048 | 313 |
| 沈阳 | 0.117 | 0.058 | 0.043 | 321 |
| 长春 | 0.099 | 0.026 | 0.039 | 340 |
| 哈尔滨 | 0.104 | 0.034 | 0.049 | 308 |
| 上海 | 0.086 | 0.051 | 0.055 | 324 |
| 南京 | 0.109 | 0.063 | 0.052 | 305 |
| 杭州 | 0.111 | 0.056 | 0.057 | 299 |
| 合肥 | 0.099 | 0.024 | 0.032 | 328 |
| 福州 | 0.072 | 0.020 | 0.049 | 344 |
| 南昌 | 0.086 | 0.056 | 0.032 | 338 |
| 济南 | 0.114 | 0.040 | 0.021 | 307 |
| 郑州 | 0.111 | 0.060 | 0.044 | 306 |
| 武汉 | 0.121 | 0.057 | 0.049 | 273 |
| 长沙 | 0.111 | 0.082 | 0.039 | 280 |
| 广州 | 0.076 | 0.054 | 0.067 | 334 |
| 南宁 | 0.066 | 0.059 | 0.035 | 353 |
| 海口 | 0.041 | 0.010 | 0.012 | 365 |
| 重庆 | 0.111 | 0.074 | 0.047 | 287 |
| 成都 | 0.123 | 0.065 | 0.049 | 301 |
| 贵阳 | 0.083 | 0.065 | 0.017 | 343 |
| 昆明 | 0.091 | 0.062 | 0.044 | 363 |
| 拉萨 | 0.062 | 0.009 | 0.026 | 363 |
| 西安 | 0.133 | 0.056 | 0.042 | 289 |
| 兰州 | 0.192 | 0.057 | 0.052 | 205 |
| 西宁 | 0.135 | 0.024 | 0.029 | 289 |
| 银川 | 0.097 | 0.048 | 0.027 | 312 |
| 乌鲁木齐 | 0.152 | 0.113 | 0.064 | 246 |

3. 聚类分析的 SPSS 过程

在数据管理窗口中定义变量名：可吸入颗粒物、二氧化硫、二氧化氮和空气质量达到以及好于二级的天数，分别用 X1、X2、X3 和 X4 来代表，之后输入原始数据，如图 12-5 所示。在数据视图（Data View）中显示的是数据集，在变量视图（Variable View）中显示的是数据集的一些性质，包括类型、长度等，用户可以自行设置。

激活“分析（Analyze）”菜单“分类（Classify）”选项中的“K-均值聚类（K-Means Cluster）”选项，如图 12-6 所示。

| 名称 | 类型 | 宽度 | 小数位数 | 标签 |
|-----------------|-----|----|------|-------------------|
| 指标 | 字符串 | 12 | 0 | |
| 可吸入颗粒物毫克立方米 | 数字 | 11 | 3 | 可吸入颗粒物(毫克/立方米) |
| 二氧化硫毫克立方米 | 数字 | 11 | 3 | 二氧化硫(毫克/立方米) |
| 二氧化氮毫克立方米 | 数字 | 11 | 3 | 二氧化氮(毫克/立方米) |
| 空气质量达到及好于二级的天数天 | 数字 | 11 | 0 | 空气质量达到及好于二级的天数(天) |

图 12-5 输入 SPSS 中的数据

然后系统会弹出如图 12-7 所示对话框，与上述 K-均值聚类 (K-Means Cluster) 过程类似，如下：

- 变量 (Variables) 栏存放分析变量，所以把 X1、X2、X3 和 X4 放入其中。
- 个案标记依据 (Label Cases by) 栏存放标识变量，所以把 City 放入其中。
- 聚类数 (Number of Clusters) 栏中的聚类数选择三类。
- 方法 (Method) 栏中选择迭代与分类 (Iterate and Classify) 方法。

选项 (Options) 框如图 12-8 所示，选择“统计量 (Statistics)”选项栏中的所有项；“缺失值 (Missing Values)”选项栏选择按列表排除个案 (Exclude Cases Listwise)。

最后，单击“K 均值聚类 (K-Means Cluster)”主对话框中的“确定”按钮，则系统会自动进行 K-Means 聚类分析。



图 12-6 “K 均值聚类 (K-Means Cluster)”菜单项



图 12-7 “K 均值聚类 (K-Means Cluster)”主选框

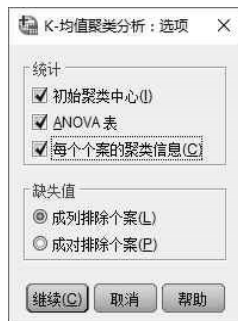


图 12-8 选项框

4. 结果分析

系统运行后会在 SPSS 输出中显示聚类分析的结果，表 12-2 为初始聚类中心，表 12-3 为迭代历史。

表 12-2 初始聚类中心

| 变量 | 聚类 | | |
|----|-----|-----|-----|
| | 1 | 2 | 3 |
| X1 | 0 | 0 | 0 |
| X2 | 0 | 0 | 0 |
| X3 | 0 | 0 | 0 |
| X4 | 234 | 365 | 300 |

上述图 12-2 中的初始聚类中心是在没有事先指定时，SPSS 会按照一定的方法从当前数据集中选取初始聚类中心。图 12-3 中的迭代过程显示第 3 次迭代后类中心就没有变化而导致迭代终止。

表 12-4 中的数据为最终聚类中心之间的距离，显然这 3 类之间的距离都比较远，故可以很好地对各个城市进行分类。

表 12-3 迭代历史

| 迭代 | 聚类中心内的更改 | | |
|----|----------|--------|-------|
| | 1 | 2 | 3 |
| 1 | 25.667 | 17.900 | 9.167 |
| 2 | 7.583 | 2.100 | .708 |
| 3 | .000 | .000 | .000 |

表 12-4 最终聚类中心间的聚类

| 聚类 | 1 | 2 | 3 |
|----|---------|---------|--------|
| 1 | | 106.750 | 60.625 |
| 2 | 106.750 | | 46.125 |
| 3 | 60.625 | 46.125 | |

表 12-5 为最终聚类结果，第一类中的城市有北京、太原、济南、武汉、长沙、重庆、兰州和乌鲁木齐；第二类中的城市有长春、合肥、福州、南昌、广州、南宁、海口、贵阳、昆明、拉萨；第三类中的城市有天津、石家庄、呼和浩特、沈阳、哈尔滨、上海、南京、杭州、郑州、成都、西安、西宁、银川。

表 12-5 聚类结果

| Case Number | City | Cluster | Distance | Case Number | City | Cluster | Distance |
|-------------|------|---------|----------|-------------|------|---------|----------|
| 1 | 北京 | 1 | 2.750 | 17 | 武汉 | 1 | 25.875 |
| 2 | 天津 | 3 | 6.125 | 18 | 长沙 | 1 | 18.875 |
| 3 | 石家庄 | 3 | 11.875 | 19 | 广州 | 2 | 11.000 |
| 4 | 太原 | 1 | 22.750 | 20 | 南宁 | 2 | 8.000 |
| 5 | 呼和浩特 | 3 | 14.125 | 21 | 海口 | 2 | 20.000 |
| 6 | 沈阳 | 3 | 22.125 | 22 | 重庆 | 1 | 11.875 |

续表

| Case Number | City | Cluster | Distance | Case Number | City | Cluster | Distance |
|-------------|------|---------|----------|-------------|------|---------|----------|
| 7 | 长春 | 2 | 5.000 | 23 | 成都 | 3 | 2.125 |
| 8 | 哈尔滨 | 3 | 9.125 | 24 | 贵阳 | 2 | 2.000 |
| 9 | 上海 | 3 | 21.000 | 25 | 昆明 | 2 | 18.000 |
| 10 | 南京 | 3 | 6.125 | 26 | 拉萨 | 2 | 18.000 |
| 11 | 杭州 | 3 | .126 | 27 | 西安 | 3 | 9.875 |
| 12 | 合肥 | 2 | 17.000 | 28 | 兰州 | 1 | 33.250 |
| 13 | 福州 | 2 | 1.000 | 29 | 西宁 | 3 | 9.875 |
| 14 | 南昌 | 2 | 7.000 | 30 | 银川 | 3 | 13.125 |
| 15 | 济南 | 1 | 8.125 | 31 | 乌鲁木齐 | 1 | 7.750 |
| 16 | 郑州 | 3 | 7.125 | | | | |

12.3 系统聚类过程

12.3.1 系统聚类简介

系统聚类法是应用最广泛的一种聚类方法，其聚类原则是相近的聚为一类，即距离最近或最相似的聚为一类。系统聚类的方法可以用于 Q 型聚类，也可以用于 R 型聚类。系统聚类法优点是既可以对观测量（样品）也可以对变量进行聚类，既可以是连续变量也可以是分类变量，提供的距离计算方法和结果显示方法也很丰富。

系统聚类法根据分析的过程不同又可以分为凝聚法和分解法两种。对于系统聚类而言，无须事先确定聚类的个数，系统会自动地将所有观测纳入计算过程中。

12.3.2 SPSS 系统聚类设置

下面将详细叙述 SPSS 中系统聚类分析的设置用法等，打开数据集，选择菜单“分析（Analyze） 分类（Classify） 系统聚类（Hierarchical Cluster）”，则系统弹出“系统聚类法”的主对话框，如图 12-9 所示，下面介绍系统聚类法的各个选项含义。

1. 主面板框的设置

主面板框如图 12-9 所示，其中的各个设置的含义如下。

- 变量（Variables）：待分析的数值型变量。
- 个案标注依据（Label Cases by）：样品的标签变量。
- 聚类（Cluster）：指定聚类分析的类型，其中个案（Cases）表示进行对观测记录的聚类分析，变量（Variables）表示进行对观测变量的聚类分析。
- 显示（Display）：聚类分析的输出内容，包括统计量（Statistics）和图（Plots）反映

聚类过程的树形图、冰状图等信息。

- 统计量 (Statistics): 统计量的输出设置框。
- 图 (Plots): 聚类图形输出设置框。
- 方法 (Method): 聚类方法, 包括类间平均法、类内平均法、最邻近距离法、中心法等。
- 保存 (Save): 结果保存设置框。

2. 统计量 (Statistics) 设置

在图 12-9 的主面板框中, 单击“统计量 (Statistics)”按钮, 则系统弹出如图 12-10 所示的统计量 (Statistics) 设置框, 其中各选项含义如下。

- 集中计划 (Agglomeration Schedule): 输出聚类的过程表, 其中包括每一步被合并的类或者观测量, 以及它们之间的距离和新生成类的信息等。
- 近似值矩阵 (Proximity Matrix): 输出各项之间的距离矩阵或者相似度矩阵。
- 聚类成员 (Cluster Membership): 设置类成员的表的输出格式有三个选择, 无 (None) 表示不显示类成员表, 单个解 (Single Solution) 表示输出指定聚类个数时的类成员表, 其右侧输入框指定聚类个数, 其值大于 1, 且小于等于聚类观测的个数和变量的个数; 解的范围 (Range of solutions) 表示输出聚类个数在某范围时的类成员表, 两个选择分别表示最小和最大的聚类个数。



图 12-9 “系统聚类法 (Hierarchical Cluster)”对话框



图 12-10 “统计量 (Statistics) 设置”对话框

3. 图 (Plots) 设置

在图 12-9 中单击“图 (Plots)”按钮, 则弹出如图 12-11 所示的图 (Plots) 选项设置对话框, 其各个选择的含义如下。

- 谱系图 (Dendrogram): 输出聚类树形图。
- 冰柱 (Icicle): 冰状图的参数设置, 有三个选择, 其中全部聚类 (All Clusters) 表示把聚类的每一步都表现在图中, 这样可以看到聚类的全过程; 指定范围的聚类

(Specified range of clusters) 表示要显示的聚类个数范围, 其有三个参数设置, 全是正整数, 开始聚类 (Start) 表示要显示的起始聚类步数, 停止聚类 (Stop) 表示要显示的终止聚类步数, 排序标准 (By) 表示要连续显示的两步聚类步骤之间的步增数量; 无 (None): 不生成冰状图。

- 方向 (Orientation): 设置冰状图的显示方向, 包括垂直方向 (Vertical) 和水平方向 (Horizontal)。

4. 方法 (Method) 设置

在图 12-9 中单击“方法 (Method)”按钮, 则弹出如图 12-12 所示的“方法 (Method) 设置”对话框, 其各个选项含义如下。

- 聚类方法 (Cluster Method): 指定聚类的方法, SPSS19.0 中提供了 7 种聚类选择。
- 测量标准 (Measure): 指定计算距离的公式, 首先是数据类型区间 (Interval)、计数 (Counts) 和二元 (Binary), 然后选择计算距离的公式。
- 转换值 (Transform Values): 设置对观测量或者变量进行标准化的参数, 其中标准化 (Standardize) 表示标准化的方法。
- 转换测量 (Transform Measure): 设置对距离测度的技术结果进行转化的方法。



图 12-11 “图 (Plots) 选项设置”对话框



图 12-12 “方法 (Method) 设置”对话框

单击“聚类方法 (Cluster Method)”选项则有以下下拉的列表, 表中即是所要选择的 7 种聚类方法, 如图 12-13 所示, 7 种方法分别介绍如下。

- 组间联接 (Between-groups Linkage): 合并两类时以两类里所有两两配对观测的平均距离最小为依据, 且配对的两个观测属于不同的类别。
- 组内联接 (Within-groups Linkage): 类内联接法, 两类合并时以所有两两配对观测的平均距离最小为依据。
- 最近邻元素 (Nearest Neighbor): 首先合并最近或者最相似的两个观测, 然后利用两个类别中的最近点之间的距离代表两个类之间的距离。
- 最远邻元素 (Furthest Neighbor): 首先合并最近或者最相似的两个观测, 然后利用

两个类别中的最远点之间的距离代表两个类之间的距离。

- 质心聚类法 (Centroid Clustering): 先计算各个类别里所有变量的均值, 再以这些均值之间的距离代表类别中间的距离。
- 中位数聚类法 (Median Clustering): 首先计算两个类之间所有配对观测的距离, 然后取这些距离的中位数代表两个类之间的距离。
- 瓦尔德法 (Ward's Method): 离差平方和法。

5. 保存 (Save) 设置

在图 12-9 中单击“保存 (Save)”按钮, 则弹出如图 12-14 所示的“保存 (Save) 设置”对话框, 对话框中“聚类成员 (Cluster Membership)”选项有三个选择。

- 无 (None): 不保存任何结果。
- 单个解 (Single Solution): 保存指定聚类个数时的分类结果, 在右侧的输入框指定聚类个数, 其值大于 1, 且小于等于参与聚类的观测个数和变量个数。
- 解的范围 (Range of Solutions): 保存聚类个数在某个范围时的分类结果, 包括最小和最大的聚类个数。



图 12-13 聚类方法 (Cluster Method) 的方法列表



图 12-14 “保存 (Save) 设置”对话框

12.3.3 实例分析



起始文件——附带光盘“PROGRAM\Ch12\实例 12-2\数据集 ch1202.xls”

动画演示——附带光盘“AVI\Ch12\12-2.avi”

1. 数据来源及说明

本案例数据来源于国家统计局网站, 数据集为我国 35 所城市 2003 年主要经济指标, 包括年末总人口数、地区生产总值 (GDP)、限额以上工业总产值、客运总量、货运总量、地方财政预算内收入、固定资产投资总额、城乡居民储蓄年末余额、在岗职工平均人数和在岗职工工资总额。

2. 分析要求

以中国大陆 35 个主要城市为研究对象,选取反映地区经济综合竞争力的 10 项重要指标,通过对原始数据的采集处理,运用 SPSS 中的层次聚类分析方法对地区差异进行研究分析,寻找各个地区的差异等。由于数据集比较大请读者自行参考光盘中的数据集 ch1202.xls。

3. SPSS 分析过程

由于原始数据的不同变量之间存在不同的量纲,不同数量级的情况,所以,为了使各个变量更具有可比性,使数据得以在更加平等的条件下进行分析处理,在进行聚类分析之前有必要对原始数据进行标准化处理,处理后的数据所得到的分析结果更具有可信性。通常情况下 SPSS 聚类分析中 Proximity 过程会先根据数据特性对原始数据进行标准化处理。

首先,把原始数据集导入到 SPSS 中,然后选择“分析 (Analyze)”菜单中“分类 (Classify)”选项的“系统聚类 (Hierarchical Cluster)”选项,单击则进入层次聚类的主对话框,然后把各种变量导入到各自变量框中,如图 12-15 所示。

- 变量 (Variables) 栏存放分析变量,所以把 10 种指标变量放入其中。
- 标注个案 (Label Cases by) 栏存放标识变量,所以把“城市”标签放入其中。
- 分群 (Clusters) 栏选择对个案进行聚类。
- 输出 (Display) 栏全选。
- 图 (Plots) 选项栏中选择谱系图选项。
- 方法 (Method) 栏中选择系统默认的组间联接 (Between-groups Linkage) 聚类方法。

上述设置完成后,单击图 12-15 中的“确定”按钮,则系统会自动完成层次聚类分析过程。

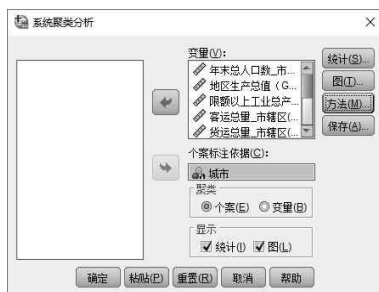


图 12-15 “层次聚类”对话框

4. 结果分析

系统完成分析以后首先得到的是处理数据集的基本信息,参见表 12-6。案例中共有 35 个样本进入聚类分析中,无缺失值。

表 12-6 数据基本信息

| 案例 | | | | | |
|----|-------|----|-----|----|-------|
| 有效 | | 缺失 | | 总计 | |
| N | 百分比 | N | 百分比 | N | 百分比 |
| 35 | 100.0 | 0 | 0 | 35 | 100.0 |

然后是聚类的凝聚过程表,此表显示此聚类方法合并过程,第一步第 4 个城市和第 30 个城市合并,其相关系数为最大 9.313E10。然后依次类推,参见表 12-7。

表 12-7 凝聚过程表

| 阶 | 群集组合 | | 系数 | 首次出现阶群集 | | 下一阶 |
|---|------|------|----------|---------|------|-----|
| | 群集 1 | 群集 2 | | 群集 1 | 群集 2 | |
| 1 | 4 | 30 | 9.313E10 | 0 | 0 | 10 |

续表

| 阶 | 群集组合 | | 系数 | 首次出现阶群集 | | 下一阶 |
|----|------|------|----------|---------|------|-----|
| | 群集 1 | 群集 2 | | 群集 1 | 群集 2 | |
| 2 | 17 | 35 | 3.461E11 | 0 | 0 | 6 |
| 3 | 5 | 26 | 5.051E11 | 0 | 0 | 7 |
| 4 | 33 | 34 | 5.591E11 | 0 | 0 | 11 |
| 5 | 3 | 15 | 5.944E11 | 0 | 0 | 10 |
| 6 | 17 | 29 | 1.018E12 | 2 | 0 | 9 |
| 7 | 5 | 25 | 1.028E12 | 3 | 0 | 11 |
| 8 | 14 | 32 | 1.986E12 | 0 | 0 | 9 |
| 9 | 14 | 17 | 2.737E12 | 8 | 6 | 19 |
| 10 | 3 | 4 | 2.984E12 | 5 | 1 | 16 |
| 11 | 5 | 33 | 3.131E12 | 7 | 4 | 24 |
| 12 | 21 | 27 | 3.778E12 | 0 | 0 | 20 |
| 13 | 28 | 31 | 4.146E12 | 0 | 0 | 18 |
| 14 | 13 | 19 | 5.599E12 | 0 | 0 | 17 |
| 15 | 20 | 22 | 5.651E12 | 0 | 0 | 16 |
| 16 | 3 | 20 | 7.312E12 | 10 | 15 | 19 |
| 17 | 8 | 13 | 8.148E12 | 0 | 14 | 25 |
| 18 | 9 | 28 | 9.589E12 | 0 | 13 | 26 |
| 19 | 3 | 14 | 9.647E12 | 16 | 9 | 24 |
| 20 | 7 | 21 | 1.124E13 | 0 | 12 | 23 |
| 21 | 16 | 18 | 1.506E13 | 0 | 0 | 25 |
| 22 | 11 | 12 | 1.685E13 | 0 | 0 | 29 |
| 23 | 6 | 7 | 1.688E13 | 0 | 20 | 26 |
| 24 | 3 | 5 | 2.043E13 | 19 | 11 | 28 |
| 25 | 8 | 16 | 2.787E13 | 17 | 21 | 27 |
| 26 | 6 | 9 | 5.282E13 | 23 | 18 | 27 |
| 27 | 6 | 8 | 7.517E13 | 26 | 25 | 28 |
| 28 | 3 | 6 | 1.344E14 | 24 | 27 | 33 |
| 29 | 2 | 11 | 1.894E14 | 0 | 22 | 31 |
| 30 | 1 | 23 | 4.314E14 | 0 | 0 | 32 |
| 31 | 2 | 24 | 6.579E14 | 29 | 0 | 32 |
| 32 | 1 | 2 | 1.099E15 | 30 | 31 | 33 |
| 33 | 1 | 3 | 1.610E15 | 32 | 28 | 34 |
| 34 | 1 | 10 | 1.107E16 | 33 | 0 | 0 |

最后的聚类结果如图 12-16 所示,当选择标尺为 5 时,则样本分为 4 类,上海单独分为一类,北京、广州和深圳分为一类,南京、杭州和天津分为一类,其余 27 个城市全部分为一类。当选择标尺为 10 时,则分为 3 类,还是一样上海为一类,北京、广州和深圳分为一类,南京、杭州、天津则和其余 27 个城市全部分为一类。其分类结果和我国现实的经济地理情况一致。由图 12-16 可以看出,上海、北京、广州、深圳等沿海城市的经济综合竞争力很强,而国内中西部地区的城市竞争力则明显较弱,聚类结果比较符合客观事实。

综上所述,可以认为研究结果基本符合近年来我国经济综合竞争力的实际情况,由于我国各个城市的经济发展条件、资源禀赋和要素不一,经济结构(尤其是产业结构)不同,经济发展状况差异较大,因此,对于不同经济综合竞争力水平及发展阶段上的地区来说,进一步提高竞争力的发展重点和发展方向也不应完全一致,应立足各个地区的区情,分类指导经济活动,鼓励和引导发挥区域比较优势、区域资源优势,制定切实可行的重点发展产业和行业,构建和培育各具特色的综合发展体系和发展方向,落实科学发展观,全面提高经济综合竞争力,实现可持续发展。

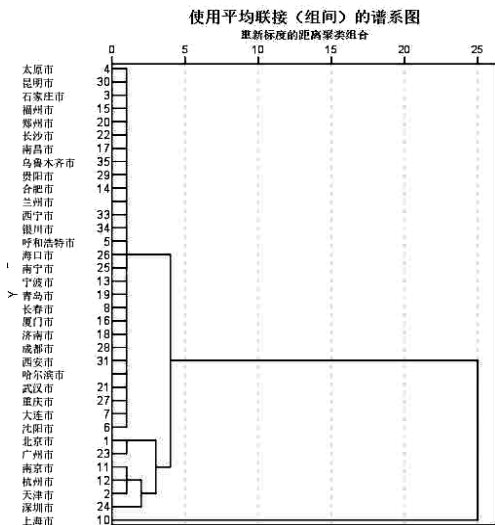


图 12-16 层次聚类结果

12.4 二阶聚类分析

12.4.1 二阶聚类简介

二阶聚类法是一种探索性的聚类方法,是随着人工智能发展起来的智能聚类方法中的一种。用于解决海量数据或具有复杂类别结构的聚类分析问题。二阶聚类法特点如下。

- 同时处理离散变量和连续变量的能力。
- 自动选择聚类数。
- 通过预先选取样本中的部分数据构建聚类模型。
- 可以处理超大样本量的数据。

二阶聚类法的基本原理是第一步:预聚类。对记录进行初始的归类,用户自定义最大

类别数。通过构建和修改特征树 (CT Free) 完成。第二步：正式聚类。对第一步完成的初步聚类进行再聚类并确定最终的聚类方案，系统根据一定的统计标准确定聚类的类别数目。以后，可以通过传统的聚类方法进行聚类 (SPSS 中采用合并型系统聚类法)。

12.4.2 SPSS 二阶聚类的设置

打开数据集，在选择菜单“分析 (Analyze) 分类 (Classify) 二阶聚类 (TwoStep Cluster)”，则系统弹出两阶段聚类法的主对话框，如图 12-17 所示，下面介绍二阶聚类法的各个选项含义。

1. 主面板框的设置

主面板框如图 12-17 所示，其中的各个设置的含义如下。

- 分类变量 (Categorical Variables)：待分析的分类变量。
- 连续变量 (Continuous Variables)：待分析的连续变量。
- 距离测量 (Distance Measure)：类别之间的距离定义选择，包括两种方法，即对数似然距离 (Log-likelihood) 和欧氏距离 (Euclidean)。
- 聚类数目 (Number of Clusters)：指定聚类分析的个数，有两种方法，包括自动确定 (Determine automatically) 和指定固定值 (Specify fixed)。
- 连续变量计数 (Count of Continuous Variables)：显示对连续变量进行标准化处理的个数统计信息，其中待标准化的计数 (To be Standardize) 显示的是要进行标准化处理的连续变量个数；假定标准化的计数 (Assumed Standardize) 显示的是不需要进行标准化的连续变量个数。
- 聚类准则 (Clustering Criterion)：指定自动聚类算法中确定最优聚类个数的准则，包括两种方法即施瓦兹贝叶斯准则 (Bayesian Information Criterion) 和赤池信息准则 (Akaike Information Criterion)。

2. 选项 (Option) 设置

由图 12-17 的主面板框，单击“选项 (Option)”按钮，则系统弹出如图 12-18 所示的“选项设置”对话框，如果再单击“高级 (Advanced)”按钮，则展开高级设置选项，其中各选项含义如下所述。



图 12-17 “二阶聚类法”主对话框



图 12-18 “选项设置”对话框

- 离群值处理 (Outlier Treatment): 设置对异常值的处理方式, 使用噪声处理 (Use Noise Handling) 表示当 CF 树长满后把稀疏节点合并为一个单独的“噪声”节点, 然后重新执行 CF 树生长过程; 百分比 (Percentage) 表示比例的临界值, 系统默认为 25%。
- 内存分配 (Memory Allocation): 设置聚类过程中所使用的内存。
- 连续变量的标准化 (Standardization of Continuous Variables): 设置对连续变量的标准化准则, 默认情况下所有变量自动选入假设标准化的变量 (To be Standardized) 表示对变量进行标准化处理。
- 树调节准则 (CF Tree Tuning Criteria): 设置决策树的调整准则, 有 3 个待定设置参数, 其中初始距离更改阈值 (Initial Distance Change Threshold) 表示指定 CF 树生长的初始临界值, 默认值为 0; 每个叶节点的最大分支数 [Maximum Branches (per leaf node)] 表示指定单个节点能够拥有的最多子节点个数, 默认值为 8; 最大树深度 (Maximum Tree Depth) 表示指定 CF 树的最大深度, 默认值为 3。
- 可能的最大节点数 (Maximum Number of Node Possible): 显示当前过程可能产生的最大节点个数。
- 聚类模型更新 (Cluster Model Update): 设置关于引入和更新旧模型的选项。

3. 输出 (Output) 设置

在图 12-17 中单击“输出 (Output)”按钮, 则弹出如图 12-19 所示的“输出 (Output) 选项”对话框, 其各个选项含义如下。

- 输出: 选择需要评估的字段。
- 工作数据文件 (Working Data File): 指定在当前数据集中保存哪些结果, 创建聚类成员变量 (Create Cluster Membership Variable) 表示保存最终的聚类结果。
- XML 文件 (XML Files): 以 XML 格式导出最终的聚类模型和 CF 树。



图 12-19 “输出 (Output) 选项”对话框

12.4.3 实例分析



起始文件

——附带光盘 “PROGRAM\Ch12\实例 12-3\数据集 ch1203.xls”



动画演示

——附带光盘“AVI\Ch12\12-3.avi”

改革开放以来,我国农村经济持续快速增长,农村居民收入与消费水平不断提高,但是自 20 世纪 90 年代以来,农民收入开始缓慢增长,处于持续低迷状态,严重影响了农民的消费需求,需求降低已成为阻碍经济进一步发展、走出当前全球经济危机的突出问题。“十一五”时期经济社会发展的主要目标是,把建设社会主义新农村作为中国现代化进程中的重大历史任务之一。深入分析农村收入和消费增长问题,积极寻求有效对策,扩大农村消费,对缓解国际金融危机对我国的影响已经成为当前及今后一个时期整个经济工作的重中之重。因此,本案例采用二阶聚类法对全国农村人均收入与消费支出水平及其关系进行量化分析研究。

1. 数据来源及说明

本案例数据来源于国家统计局网站,数据集为 2006 年农村居民家庭人均纯收入和农村居民家庭人均消费情况两组指标。其中农村居民家庭人均纯收入包括工资性收入、家庭经营纯收入、财产性收入和转移性收入指标;农村居民家庭人均消费包括食品、衣着、居住、家庭设备及服务、交通和通信、文教娱乐服务和医疗保健指标。数据集见光盘中 Ch12\实例 12-3\ch1203.xls 文件。

2. 分析要求

采用二阶聚类分析法对数据进行聚类分析,首先是预聚类,对原始数据进行分析,构建聚类特征树。然后是正式聚类,利用特征树进行系统聚类分析。然后来分析不同变量在各个类别之中的重要性,以及各个类别中的收入支出类型特征。

3. SPSS 分析过程

首先,导入原始数据集,同其他聚类分析方法一样,然后选择“分析(Analyze)”菜单中“分类(Classify)”选项的“二阶段聚类(TwoStep Cluster)”选项,单击则进入“二阶聚类”的主对话框,然后把各种变量导入到各自变量框中,如图 12-20 所示。

- 分类变量(Categorical Variables)栏存放分类标识变量,所以把“各省市”放入其中。
- 连续变量(Continuous Variables)栏存放连续变量,所以把“工资性收入”等 12 个变量放入其中。
- 聚类数量(Number of Clusters)栏中选择指定固定数(Specify Fixed),设置为 4 类。

最后,单击图 12-20 中的“确定”按钮,则系统自动进行二阶聚类分析。



图 12-20 “二阶聚类”主对话框

4. 结果分析

二阶聚类分析运行以后, 会在 SPSS 输出窗口中输出大量结果, 下面将对最重要的输出表格进行分析。首先是聚类结果的基本统计信息, 参见表 12-8, 聚类分布表格中给出了最终的分类结果, 可见第 3 类的观测数最多, 其次是 2 和 4 类, 最后的是第 1 类。

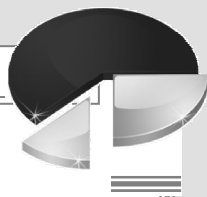
其次是连续变量的基本统计信息, 图 12-21 中给出了关于连续变量的均值和标准差信息。第 1 类中的工资性收入均值为 5102.83 元, 在 4 个类别中最大。其他 3 个类别在此变量上明显要小很多。

表 12-8 基本统计信息

| | | N | 组合/% | 总计/% |
|----|----|----|--------|--------|
| 聚类 | 1 | 3 | 9.7% | 9.7% |
| | 2 | 8 | 25.8% | 25.8% |
| | 3 | 12 | 38.7% | 38.7% |
| | 4 | 8 | 25.8% | 25.8% |
| | 组合 | 31 | 100.0% | 100.0% |
| 总计 | | 31 | | 100.0% |

| | | 聚类 | | | | |
|------------|-----|----------|---------|---------|----------|----------|
| | | 1 | 2 | 3 | 4 | 组合 |
| 工资性收入 | 平均值 | 5102.83 | 1011.53 | 895.66 | 1935.15 | 1600.96 |
| | 标准差 | 1556.191 | 474.161 | 312.341 | 1060.765 | 1435.254 |
| 家庭经营纯收入 | 平均值 | 1936.37 | 2147.67 | 1465.42 | 2342.19 | 1913.32 |
| | 标准差 | 1158.439 | 279.217 | 196.905 | 304.828 | 536.235 |
| 财产性收入 | 平均值 | 516.20 | 63.18 | 63.69 | 147.65 | 129.01 |
| | 标准差 | 187.168 | 41.621 | 36.537 | 44.602 | 147.106 |
| 转移性收入 | 平均值 | 694.27 | 137.20 | 156.86 | 249.70 | 227.75 |
| | 标准差 | 391.606 | 46.225 | 66.286 | 74.711 | 199.310 |
| 食品 | 平均值 | 2373.80 | 1100.88 | 994.61 | 1371.11 | 1252.66 |
| | 标准差 | 587.763 | 239.555 | 137.759 | 321.879 | 479.345 |
| 衣着 | 平均值 | 412.63 | 150.54 | 135.43 | 208.75 | 185.08 |
| | 标准差 | 41.673 | 38.364 | 44.461 | 35.697 | 90.072 |
| 居住 | 平均值 | 1239.83 | 416.36 | 352.83 | 532.23 | 501.36 |
| | 标准差 | 400.692 | 97.826 | 74.200 | 124.818 | 290.053 |
| 家庭设备及服务 | 平均值 | 357.57 | 104.01 | 100.31 | 138.98 | 136.14 |
| | 标准差 | 107.170 | 22.777 | 17.989 | 35.386 | 83.615 |
| 交通和通讯 | 平均值 | 714.27 | 241.92 | 195.72 | 391.06 | 308.24 |
| | 标准差 | 59.229 | 28.065 | 47.895 | 61.774 | 163.406 |
| 文教、娱乐用品及服务 | 平均值 | 831.87 | 257.59 | 200.61 | 375.68 | 321.58 |
| | 标准差 | 94.744 | 58.170 | 78.767 | 77.336 | 197.037 |
| 医疗保健 | 平均值 | 528.20 | 173.66 | 143.70 | 229.20 | 210.71 |
| | 标准差 | 61.027 | 42.259 | 43.855 | 35.813 | 118.333 |
| 其他商品及服务 | 平均值 | 137.70 | 58.75 | 41.45 | 82.64 | 65.86 |
| | 标准差 | 34.002 | 16.138 | 7.937 | 22.489 | 33.495 |

图 12-21 连续变量基本统计信息



第 13 章 判别分析

判别分析 (Discriminant Analysis), 又称“分辨法”, 属于分类方法的一种, 分类的对象要求事先要有明确的类别空间, 这一点与聚类分析迥然不同。它是在分类确定的条件下, 根据某一研究对象的各种特征值判别其类型归属问题的一种多变量统计分析方法。其基本原理是按照一定的判别准则, 建立一个或多个判别函数, 用研究对象的大量资料确定判别函数中的待定系数, 并计算判别指标。据此即可确定某一样本属于何类。

判别分析有多种方法, 例如, 最大似然法、Fisher 判别分析法、Bayes 判别分析法、逐步判别分析法等, 距离判别和典型判别对数据分布无严格要求, 而 Bayes 判别分析法则要求数据服从多元正态分布。判别分析在气候分类、农业区划、土地类型划分中有着广泛的应用。不同的判别分析方法有其特定的适应条件, 掌握各种方法的适用条件是保证正确分析结果可靠性的重要条件。



本讲内容

- 判别分析的基本原理
- 一般判别分析
- 逐步判别分析

13.1 判别分析的基本原理

判别分析是一种判别个体所隶属的群体的统计分析手段, 是根据已知对象的某些观测指标和所属类别来判断未知对象所属类别的一种统计学方法。其作用表现在, 当描述研究对象的性质特征不全或不能从直接测量数据确定研究对象所属类别时, 可以通过判别分析对其进行归类。下面将简要讲述为什么要进行判别分析。

13.1.1 判别分析简介

判别分析是判别样品所属类型的一种统计方法, 其应用之广泛可与回归分析相媲美。

在生产、科研和日常生活中经常需要根据观测到的数据资料, 对所研究的对象进行分类。例如, 银行在贷款给顾客时, 通常都会根据顾客的基本资料, 如学历、收入、借贷记

录等,将顾客区分为具有信用之顾客与不具有信用之顾客两种,并且当有新的顾客进来时,也可以按照同样准则将新顾客的资料与这些已经存在的资料做一比较,看是否应该借钱给这位新的顾客;在经济学中,根据人均国民收入、人均工农业产值、人均消费水平等多种指标来判定一个国家的经济发展程度所属类型;在市场预测中,根据以往调查所得的种种指标,判别下季度产品是畅销、平常或滞销;在生物物种研究时,例如,有一些昆虫的性别很难看出,只有通过解剖才能判别,但是雄性和雌性昆虫在若干体表度量上有些综合的差异,于是统计学家就根据已知雌雄的昆虫体表度量(这些用作度量的变量称为预测变量)得到一个标准,并且利用这个标准来判别其他未知性别的昆虫,这样的判别虽然不能保证百分之百准确,但至少大部分判别都是对的,而且用不着杀死昆虫来进行判别;在天气预报中,可以根据某些气象资料来判断近期的天气变化,需要将这些气象资料同某些典型的天气变化规律进行对照,判断最可能的情况;在体育运动中,判别某游泳运动员“苗子”是适合练蛙泳、仰泳,还是自由泳等;在医疗诊断中,根据某人多种体验指标(如体温、血压、白血球等)来判别此人是有病还是无病。总之,在实际问题中需要判别的问题几乎到处可见。

判别分析与聚类分析不同。判别分析是在已知研究对象分成若干类型(或组别)并已取得各种类型的一批已知样品的观测数据,在此基础上根据某些准则建立判别式,然后对未知类型的样品进行判别分类。对于聚类分析来说,一批给定样品要划分的类型事先并不知道,正需要通过聚类分析来确定类型。

正因为如此,判别分析和聚类分析往往联合起来使用,例如,判别分析是要求先知道各类总体情况才能判断新样品的归类,当总体分类不清楚时,可先用聚类分析对原来的一批样品进行分类,然后再用判别分析建立判别式以对新样品进行判别。

判别分析内容很丰富,方法很多。判别分析按判别的组数来区分,有两组判别分析和多组判别分析;按区分不同总体的所用的数学模型来分,有线性判别和非线性判别;按判别时所处理的变量方法不同,有逐步判别和序贯判别等。判别分析可以从不同角度提出问题,因此有不同的判别准则,如马氏距离最小准则、Fisher 准则、平均损失最小准则、最小平方准则、最大似然准则、最大概率准则等,按判别准则的不同又提出多种判别方法。本章仅介绍四种常用的判别方法,即距离判别法、Fisher 判别法、Bayes 判别法和逐步判别法。

判别分析对气候分类、农业区划、医学研究、信用风险管理等课题的研究有非常重要的作用。在利用 SPSS 软件系统进行判别分析之前,先了解有关判别分析的基本概念、基本原理,以及判别分析过程等。

13.1.2 判别分析的数学模型与判别方法

已知某食物有 K 个状态(K 个类),这 K 个状态可以看作 K 个总体 G_1, G_2, \dots, G_K , 该事物的特性可以由 P 个指标 X_1, X_2, \dots, X_p 来刻画,并在分析前已经观察到了总体 G_1, G_2, \dots, G_K 的 n_1, n_2, \dots, n_K 个样品。 G_i 的分布参见表 13-1, 其中 $i = 1, 2, 3, \dots, K$, $n = n_1 + n_2 + n_3 + \dots + n_K$ 。

判别分析就是根据以上观测数据,依据某种判别标准建立一个判别函数,并根据该函

数对新样品进行判别归类。判别分析的任务是根据已掌握的样本资料，建立判别函数，进而对给定的新观察，判断它来自哪一个总体。例如，企业是否陷入财务困境；上市公司是 ST 或 PT 类公司还是非 ST 或非 PT 类等。

上述的判别函数是一个判别准则，那么该准则是如何得到的呢？判别分析的目的就是判断给定的新观测属于哪一个总体，故关键是判断的依据。下面将详细介绍距离判别法、Fisher 判别法和 Bayes 判别法。

1. 距离判别法

距离判别法的基本思想是根据已知分类的数据，分别计算各类的重心即分组（类）的均值，判别准则是对任给的一次观测，若它与第 i 类的重心距离最近，就认为它来自第 i 类。这里要注意的是距离判别法，对各类（或总体）的分布，并无特定的要求。下面详细讲述两个总体的距离判别法，多个总体的距离判别法类似于两个总体的距离判别法，可以很容易地推广得到。

设有两个总体（或称两类） G_1 、 G_2 ，从第一个总体中抽取 n_1 个样品，从第二个总体中抽取 n_2 个样品，每个样品测量 p 个指标参见表 13-1。

表 13-1 G_i 总体分布

| 变量 样品 | G_i 总体 | | | |
|-------------|-------------------|-------------------|----------|-------------------|
| | x_1 | x_2 | \cdots | x_p |
| $x_1^{(i)}$ | $x_{11}^{(i)}$ | $x_{12}^{(i)}$ | \cdots | $x_{1p}^{(i)}$ |
| $x_2^{(i)}$ | $x_{21}^{(i)}$ | $x_{22}^{(i)}$ | \cdots | $x_{2p}^{(i)}$ |
| \vdots | \vdots | \vdots | \cdots | \vdots |
| $x_n^{(i)}$ | $x_{n1}^{(i)}$ | $x_{n2}^{(i)}$ | \cdots | $x_{np}^{(i)}$ |
| 均值 | $\bar{x}_1^{(i)}$ | $\bar{x}_2^{(i)}$ | \cdots | $\bar{x}_p^{(i)}$ |

今任取一个样品，实测指标值为 $X = (x_1, \cdots, x_p)'$ ，问 X 应判归为哪一类？

首先计算 X 到 G_1 、 G_2 总体的距离，分别记为 $D(X, G_1)$ 和 $D(X, G_2)$ ，按距离最近准则判别归类，则可写成

$$\begin{cases} X \in G_1, & D(X, G_1) < D(X, G_2) \\ X \in G_2, & D(X, G_1) > D(X, G_2) \\ \text{待判,} & D(X, G_1) = D(X, G_2) \end{cases}$$

记 $\bar{X}^{(i)} = (\bar{x}_1^{(i)}, \cdots, \bar{x}_p^{(i)})'$ ， $i = 1, 2$ ，如果距离定义采用欧氏距离，则可计算出

$$D(X, G_1) = \sqrt{(X - \bar{X}^{(1)})'(X - \bar{X}^{(1)})} = \sqrt{\sum_{a=1}^p (x_a - \bar{x}_a^{(1)})^2}$$

$$D(X, G_2) = \sqrt{(X - \bar{X}^{(2)})'(X - \bar{X}^{(2)})} = \sqrt{\sum_{a=1}^p (x_a - \bar{x}_a^{(2)})^2}$$

然后比较 $D(X, G_1)$ 和 $D(X, G_2)$ 大小，按距离最近准则判别归类。

2. 费希尔 (Fisher) 判别法

费希尔 (Fisher) 判别法是 1936 年提出来的, 其思想是投影, 把多维问题简化为一维问题来处理。该法对总体的分布并未提出特定的要求。

(1) 基本思想

从两个总体中抽取具有 p 个指标的样品观测数据, 借助方差分析的思想造一个判别函数或称判别式为 $y = c_1x_1 + c_2x_2 + \cdots + c_px_p$, 其中系数 c_1, c_2, \dots, c_p 确定的原则是使两组间的区别最大, 而使每个组内部的离差最小。有了判别式后, 对于一个新的样品, 将它的 p 个指标值代入判别式中求出 y 值, 然后与判别临界值 (或称分界点, 后面给出) 进行比较, 就可以判别它应属于哪一个总体。

(2) 判别函数的导出

假设有两个总体 G_1, G_2 , 从第一个总体中抽取 n_1 个样品, 从第二个总体中抽取 n_2 个样品, 每个样品观测 p 个指标。

假设新建立的判别式为 $y = c_1x_1 + c_2x_2 + \cdots + c_px_p$, 今将属于不同两总体的样品观测值代入判别式中, 得

$$\begin{aligned} y_i^{(1)} &= c_1x_{i1}^{(1)} + c_2x_{i2}^{(1)} + \cdots + c_px_{ip}^{(1)}, \quad i=1, \dots, n_1 \\ y_i^{(2)} &= c_1x_{i1}^{(2)} + c_2x_{i2}^{(2)} + \cdots + c_px_{ip}^{(2)}, \quad i=1, \dots, n_2 \end{aligned}$$

对上边两式分别左右相加, 再乘以相应的样品个数, 则有第一组样品的“重心”为

$$\bar{y}^{(1)} = \sum_{k=1}^p c_k \bar{x}_k^{(1)}$$

第二组样品的“重心”为

$$\bar{y}^{(2)} = \sum_{k=1}^p c_k \bar{x}_k^{(2)}$$

为了使判别函数能够很好地区别来自不同总体的样品, 来自不同总体的两个平均值 $\bar{y}^{(1)}, \bar{y}^{(2)}$ 相差越大越好。

对于来自第一个总体的 $\bar{y}_i^{(1)}, i=1, \dots, n_1$, 要求它们的离差平方和 $\sum_{i=1}^{n_1} (y_i^{(1)} - \bar{y}^{(1)})^2$ 越小越好, 同样也要求 $\sum_{i=1}^{n_2} (y_i^{(2)} - \bar{y}^{(2)})^2$ 越小越好。

由上可知, 就是使得

$$I = \frac{(\bar{y}^{(1)} - \bar{y}^{(2)})^2}{\sum_{i=1}^{n_1} (y_i^{(1)} - \bar{y}^{(1)})^2 + \sum_{i=1}^{n_2} (y_i^{(2)} - \bar{y}^{(2)})^2}$$

越大越好。记 $Q = Q(c_1, c_2, \dots, c_p) = (\bar{y}^{(1)} - \bar{y}^{(2)})^2$ 为两组间离差, 则

$$F = F(c_1, c_2, \dots, c_p) = \sum_{i=1}^{n_1} (y_i^{(1)} - \bar{y}^{(1)})^2 + \sum_{i=1}^{n_2} (y_i^{(2)} - \bar{y}^{(2)})^2$$

为两组内的离差，则

$$I = \frac{Q}{F}$$

再利用微积分求极值的必要条件即可求出使 I 达到最大值的 c_1, c_2, \dots, c_p 。为此将上式两边取对数，则有

$$\frac{\partial \ln I}{\partial c_k} = \frac{\partial \ln Q}{\partial c_k} - \frac{\partial \ln F}{\partial c_k} = 0$$

式中， $k=1, \dots, p$ ，所以

$$\frac{1}{Q} \cdot \frac{\partial Q}{\partial c_k} = \frac{1}{F} \cdot \frac{\partial F}{\partial c_k}$$

即

$$\frac{1}{I} \cdot \frac{\partial Q}{\partial c_k} = \frac{\partial F}{\partial c_k}$$

而

$$\begin{aligned} Q &= (\overline{y^{(1)}} - \overline{y^{(2)}})^2 = \left(\sum_{k=1}^p c_k \overline{x_k^{(1)}} - \sum_{k=1}^p c_k \overline{x_k^{(2)}} \right)^2 \\ &= \left[\sum_{k=1}^p c_k (\overline{x_k^{(1)}} - \overline{x_k^{(2)}}) \right]^2 \\ &= \left[\sum_{k=1}^p c_k d_k \right]^2 \end{aligned}$$

式中， $d_k = \overline{x_k^{(1)}} - \overline{x_k^{(2)}}$ ，所以 $\frac{\partial Q}{\partial c_k} = 2 \left(\sum_{l=1}^p c_l d_l \right) d_k$ ，而

$$F = \sum_{k=1}^p \sum_{l=1}^p c_k c_l s_{kl}$$

式中， $s_{kl} = \sum_{i=1}^{n_1} (x_{ik}^{(1)} - \overline{x_k^{(1)}})(x_{il}^{(1)} - \overline{x_l^{(1)}}) + \sum_{i=1}^{n_2} (x_{ik}^{(2)} - \overline{x_k^{(2)}})(x_{il}^{(2)} - \overline{x_l^{(2)}})$ ，所以

$$\frac{\partial F}{\partial c_k} = 2 \sum_{l=1}^p c_l s_{kl}$$

从而 $\frac{2}{I} \left(\sum_{l=1}^p c_l d_l \right) d_k = 2 \sum_{l=1}^p c_l s_{kl}$ ，即

$$\frac{1}{I} \left(\sum_{l=1}^p c_l d_l \right) d_k = \sum_{l=1}^p c_l s_{kl}$$

式中， $k=1, \dots, p$ ，令 $\beta = \frac{1}{I} \sum_{l=1}^p c_l d_l$ ， β 是常数因子，不依赖于 k ，不影响其解 c_1, \dots, c_p 之间的相对比例关系。对判别结果没有影响，所以，取 $\beta=1$ ，于是

$$\sum_{l=1}^p c_l s_{kl} = d_k$$

式中, $k=1, \dots, p$, 即

$$\begin{cases} s_{11}c_1 + s_{12}c_2 + \dots + s_{1p}c_p = d_1 \\ s_{21}c_1 + s_{22}c_2 + \dots + s_{2p}c_p = d_2 \\ \vdots \\ s_{p1}c_1 + s_{p2}c_2 + \dots + s_{pp}c_p = d_p \end{cases}$$

解上述方程组即得

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}^{-1} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_p \end{bmatrix}$$

建立判别准则还要确定判别临界值 y_0 , 在两总体先验概率相等的假设下, 取 y_0 为 $\overline{y^{(1)}}$ 与 $\overline{y^{(2)}}$ 的加权平均值, 即

$$y_0 = \frac{n_1 \overline{y^{(1)}} + n_2 \overline{y^{(2)}}}{n_1 + n_2}$$

由原始数据求得 $\overline{y^{(1)}}$ 与 $\overline{y^{(2)}}$ 满足 $\overline{y^{(1)}} > \overline{y^{(2)}}$, 则判别准则为, 对一个新样品 $X = (x_1, \dots, x_p)'$ 代入判别函数中将所得值记为 y , 若 $y > y_0$, 则判定 $X \in G_1$; 若 $y < y_0$, 则判定 $X \in G_2$ 。如果 $\overline{y^{(1)}} < \overline{y^{(2)}}$, 则建立判别准则为, 若 $y > y_0$, 则判定 $X \in G_2$; 若 $y < y_0$, 则判定 $X \in G_1$ 。

(3) 计算步骤

步骤 1: 建立判别函数。

求 $I = \frac{Q(c_1, \dots, c_p)}{F(c_1, \dots, c_p)}$ 的最大值点 c_1, c_2, \dots, c_p , 根据极值原理, 解如下方程组

$$\begin{cases} \frac{\partial \ln I}{\partial c_1} = 0 \\ \frac{\partial \ln I}{\partial c_2} = 0 \\ \vdots \\ \frac{\partial \ln I}{\partial c_p} = 0 \end{cases}$$

可得到 c_1, \dots, c_p , 写出判别函数 $y = c_1 x_1 + \dots + c_p x_p$ 。

步骤 2: 计算判别临界值 y_0 , 然后根据判别准则对新样品判别分类。

步骤 3: 检验判别效果 (当两个总体协方差矩阵相同且总体服从正态分布)。

$$H_0: Ex_a^{(1)} = \mu_1 = Ex_a^{(2)} = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

检验统计量:

$$F = \frac{(n_1 + n_2 - 2) - p + 1}{(n_1 + n_2 - 2)p} T^2 \underset{\text{(在 } H_0 \text{ 成立)}}{\sim} F(p, n_1 + n_2 - p - 1)$$

其中

$$\begin{aligned}\overline{X^{(i)}} &= (x_1^{(i)}, \dots, x_p^{(i)})' \\ T^2 &= (n_1 + n_2 - 2) \cdot \left[\sqrt{\frac{n_1 n_2}{n_1 + n_2}} (\overline{X^{(1)}} - \overline{X^{(2)}})' S^{-1} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} (\overline{X^{(1)}} - \overline{X^{(2)}}) \right] \\ S &= (s_{ij})_{p \times p}, s_{ij} = \sum_{a=1}^{n_1} (x_{ai}^{(1)} - \overline{x_i^{(1)}})(x_{aj}^{(1)} - \overline{x_j^{(1)}}) + \sum_{a=1}^{n_2} (x_{ai}^{(2)} - \overline{x_i^{(2)}})(x_{aj}^{(2)} - \overline{x_j^{(2)}})\end{aligned}$$

给定检验水平 α , 查 F 分布表, 确定临界值 F_α , 若 $F > F_\alpha$, 则 H_0 被否定, 认为判别有效。否则认为判别无效。

3. Bayes 判别法

1) Bayes 判别法的基本思想

Bayes 判别法的基本思想是根据先验概率分布求出后验概率分布, 并依据后验概率分布作出统计判别。设有 k 个总体 G_1, G_2, \dots, G_k , 它们的先验概率分别为 q_1, q_2, \dots, q_k (它们可以由经验给出也可以估出)。各总体的密度函数分别为 $f_1(x), f_2(x), \dots, f_k(x)$, 在观测到一个样品 x 的情况下, 可用著名的 Bayes 公式计算它来自第 g 总体的后验概率, 即

$$P(g/x) = \frac{q_g f_g(x)}{\sum_{i=1}^k q_i f_i(x)}$$

式中, $g=1, \dots, k$, 并且当

$$P(h/x) = \max_{1 \leq g \leq k} P(g/x)$$

时, 则判 x 来自第 h 总体。

2) 多元正态总体的 Bayes 判别法

(1) 判别函数的导出

使用 Bayes 判别法作判别分析, 首先需要知道待判总体的先验概率 q_g 和密度函数 $f_g(x)$ 。 p 元正态分布密度函数为

$$f_g(x) = (2\pi)^{-p/2} |\Sigma^{(g)}|^{-1/2} \cdot \exp \left\{ -\frac{1}{2} (x - \mu^{(g)})' \Sigma^{(g)-1} (x - \mu^{(g)}) \right\}$$

式中, $\mu^{(g)}$ 和 $\Sigma^{(g)}$ 分别是第 g 总体的均值矢量 (p 维) 和协差阵 (p 阶)。把 $f_g(x)$ 代入 $P(g/x)$ 的表达式中, 因为只关心寻找使 $P(g/x)$ 最大的 g , 而分式中的分母不论 g 为何值都是常数, 故可令

$$q_g f_g(x) \xrightarrow{g} \max$$

取对数并去掉与 g 无关的项, 记为

$$\begin{aligned}Z(g/x) &= \ln q_g - \frac{1}{2} \ln |E^{(g)}| - \frac{1}{2} (x - \mu^{(g)})' \Sigma^{(g)-1} (x - \mu^{(g)}) \\ &= \ln q_g - \frac{1}{2} \ln |E^{(g)}| - \frac{1}{2} x' \Sigma^{(g)-1} x - \frac{1}{2} \mu^{(g)'} \Sigma^{(g)-1} \mu^{(g)} + x' \Sigma^{(g)-1} \mu^{(g)}\end{aligned}$$

则问题化为

$$Z(g/x) \xrightarrow{g} \max$$

(2) 假设协方差阵相等

$Z(g/x)$ 中含有 k 个总体的协方差阵 (逆阵及行列式值), 而且对于 x 还是二次函数, 实际计算时工作量很大。如果进一步假设 k 个总体协方差阵相同, 即 $\Sigma^{(1)} = \Sigma^{(2)} = \dots = \Sigma^{(K)} = \Sigma$, 这时 $Z(g/x)$ 中 $\frac{1}{2} \ln |\Sigma^{(g)}|$ 和 $\frac{1}{2} x' \Sigma^{(g)-1} x$ 两项与 g 无关, 求最大时可以去掉, 最终得到如下形式的判别函数与判别准则 (如果协方差阵不等, 则有非线性判别函数), 即

$$\begin{cases} y(g/x) = \ln q_g - \frac{1}{2} \mu^{(g)'} \Sigma^{-1} \mu^{(g)} + x' \Sigma^{-1} \mu^{(g)} \\ y(g/x) \xrightarrow{g} \max \end{cases}$$

上式判别函数也可以写成多项式形式, 即

$$y(g/x) = \ln q_g + C_0^{(g)} + \sum_{i=1}^p C_i^{(g)} x_i$$

此处

$$\begin{aligned} C_i^{(g)} &= \sum_{j=1}^p v_{ij} \mu_j^{(g)}, \quad C_0^{(g)} = -\frac{1}{2} \sum_{i=1}^p C_i^{(g)} \mu_i^{(g)} \\ x &= (x_1, x_2, \dots, x_p)' , \quad \mu^{(g)} = (\mu_1^{(g)}, \mu_2^{(g)}, \dots, \mu_p^{(g)})' \\ \Sigma &= (v_{ij})_{p \times p}, \quad \Sigma^{-1} = (v_{ij}^{-1})_{p \times p} \end{aligned}$$

式中, $i=1, \dots, p$ 。

(3) 计算后验概率

作计算分类时, 主要根据判别式 $y(g/x)$ 的大小, 而不是后验概率 $P(g/x)$, 但是有了 $y(g/x)$ 之后, 就可以根据下式算出 $P(g/x)$, 即

$$P(g/x) = \frac{\exp\{y(g/x)\}}{\sum_{i=1}^k \exp\{y(i/x)\}}$$

因为

$$y(g/x) = \ln(q_g f_g(x)) - \Delta(x)$$

其中 $\Delta(x)$ 是 $\ln[q_g f_g(x)]$ 中与 g 无关的部分。所以

$$P(g/x) = \frac{\exp\{y(g/x)\}}{\sum_{i=1}^k \exp\{y(i/x)\}}$$

由上式可知, 使 y 为最大的 h , 其 $P(h/x)$ 必为最大, 因此, 只须把样品 x 代入判别式中, 分别计算 $y(g/x)$, $g=1, \dots, k$ 。若

$$y(g/x) = \max_{1 \leq g \leq k} \{y(g/x)\}$$

则把样品 x 归入第 h 总体。

13.2 一般判别分析

13.2.1 一般判别分析的参数设置

首先,介绍一般判别分析的 SPSS 分析设置,依次选择菜单“分析 (Analyze) 分类 (Classify) 判别式 (Discriminant)”即可弹出“判别分析”对话框,如图 13-1 所示。

在此界面上,有变量设置、保存选项设置、输出选项设置,以及分类参数设置,首先是变量设置。

1. 变量设置

如图 13-1 所示,图中最左边是变量列表框,用于存放所要分析的变量,右边依次是分组变量 (Grouping Variable) 框,可以把分类变量选入此框中;然后是自变量 (Independents) 框,用于选入自变量列表;最后的选择变量 (Selection Variable) 框,用于样本筛选。各个选项框功能具体如下。

- 分组变量 (Grouping Variable): 用于选入分类变量,此变量用于标识观测变量所属的类别。单击“定义范围 (Defining Range)”按钮,可以在对话框中设置最小值和最大值,分别指定分类变量的最小值和最大值,如图 13-2 所示。



图 13-1 “判别分析 (Discriminant)”对话框



图 13-2 定义范围 (Defining Range) 对话框

- 自变量 (Independents): 用于从左边变量列表中选入自变量用于判别分析。
- 选择变量 (Selection): 选入对观测样本进行筛选的变量。选入筛选变量后就可以激活“值 (Value)”按钮,单击此按钮,可以弹出“定义变量取值”的对话框,如图 13-3 所示,可以输入取值,当筛选变量取值为这个值时,所对应的观测记录才能用来进行判别函数的推导。

在变量框自变量 (Independents) 的下面,SPSS 还给出了两种变量选择的方法。

- 一起输出自变量 (Enter Independents Together): 表示建立包括自变量的全模型。
- 使用步进法 (Use Stepwise Method): 逐步判别方法,需要根据各个变量对判别贡献的大小进行选择。选中后,图 13-1 右上角的“方法 (Method)”按钮则会被点亮,单击此按钮,则会弹出关于逐步判别法的有关参数设置。

2. 输出设置

单击图 13-1 右上角的“统计量 (Statistics)”按钮,则会弹出如图 13-4 所示的对话

框,此对话框主要是设置一些统计量,并进行输出。各选项含义如下所述。

(1) 描述性 (Descriptive) 栏

此栏主要设置一些描述性的统计量。

- 平均值 (Means): 输出各个类别中各个自变量的均值、标准差等信息。
- 单变量 (Univariate ANOVAs): 输出方差分析结果。
- 博克斯 M: 输出协方差分析结果,检验各类别的协方差矩阵是否相等。

(2) 函数系数 (Function Coefficients) 栏

此栏主要选择判别函数系数的输出形式。



图 13-3 筛选值设定



图 13-4 “统计量 (Statistics) 设置”对话框

- 费希尔 Fisher: 直接对新样本进行判别分析的 Fisher 系数。
- 未标准化 (Unstandardized): 未经标准化处理的判别系数。

(3) 矩阵 (Matrices) 栏

此栏用于输出矩阵。

- 组内相关性 (Within-groups Correlation)
- 组内协方差 (Within-groups Covariance)
- 分组协方差 (Separate-groups Covariance)
- 总协方差 (Total Covariance)

3. 保存设置

单击如图 13-1 所示的“保存 (Save)”按钮,则会弹出如图 13-5 所示的对话框,SPSS 提供了四个设置选项,各个选项功能如下。

- 预测组成员 (Predicted Group Membership): 保存观测的预测分类结果。
- 判别得分 (Discriminant Score): 保存观测的判别得分,该分数由未标准化的判别系数乘以自变量的取值再求和而得到。
- 组成员概率 (Probabilities of Group Membership): 保存观测记录属于某一类的概率。
- 将模型信息导出到 XML 文件 (Export Model Information to XML File): 把模型信息保存到指定的 XML 文件之中,单击“浏览 (Browse)”按钮选择文件路径和名称。

4. 分类设置

单击如图 13-1 所示的“分类 (Classify)”按钮,则弹出如图 13-6 所示的对话框,各个选项栏具体功能如下。

(1) 先验概率 (Prior Probabilities) 栏

此栏用于指定先验概率。

- 所有组相等 (All Groups Equal): 各类别的先验概率相等。
- 根据组大小计算 (Compute from Group Sizes): 表示各类别的先验概率与其样本量成正比。



图 13-5 “保存 (Save) 设置”对话框



图 13-6 “分类 (Classify) 设置”对话框

(2) 显示 (Display) 栏

此栏用于设置分类结果的输出选项。

- 个案结果 (Casewise Results): 输出对单个观测量的详细分类信息。Limits cases to 复选框设置输出的范围, 若输入 n , 表示对前 n 个观测有输出。
- 摘要表 (Summary Table): 输出分类总结表, 包括正确分类的观测数目和错分观测的分类数目, 以及正确率和错误率。
- 留一分类 (Leave-one-out Classification): 输出交互校验信息, 由除去单个观测以外的其他观测导出的判别函数预测这个观测的类别, 输出如此得到的统计信息。

(3) 使用协方差矩阵 (Use Covariance Matrix) 栏

此栏设置分类时所使用的协方差矩阵。

- 组内 (Within-groups): 指定使用合并的类内协方差矩阵进行分类。
- 分组 (Separate-groups): 指定使用每个类别的协方差矩阵进行分类。

(4) 图 (Plots) 栏

- 合并组 (Combine-groups): 联合散点图。
- 分组 (Separate-groups): 多张散点图。
- 领域图 (Territorial map): 边界图。

13.2.2 实例分析



结果文件

——附带光盘 “PROGRAM\CH13\实例 13-1” 文件夹



动画演示

——附带光盘 “AVI\实例 13-1.avi” 文件

本例讨论的上市公司的类型主要划分为 ST 公司和非 ST 公司, 案例分析就是要得出给定公司是 ST 公司还是非 ST 公司的判别函数。

现代财务理论认为,财务比率指标能将资产负债表、利润表和现金流量表有机结合起来,而且在一定程度上能消除企业规模影响。我们选择反映企业偿债能力、运营能力、盈利能力、成长能力和现金流量等方面的常用财务指标,具体参见表 13-2。

表 13-2 财务指标

| 指 标 | 符 号 | 指 标 | 符 号 |
|--------------|----------|-----------|----------|
| 流动比率 | X_1 | 现金比率 | X_2 |
| 流动负债经营活动现金流比 | X_3 | 资产负债率 | X_4 |
| 应收账款周转率 | X_5 | 存货周转率 | X_6 |
| 流动资产周转率 | X_7 | 总资产周转率 | X_8 |
| 主营业务利润率 | X_9 | 净资产收益率 | X_{10} |
| 每股收益 | X_{11} | 主营业务收入增长率 | X_{12} |
| 净利润增长率 | X_{13} | 总资产扩张率 | X_{14} |
| 每股营业现金流量 | X_{15} | 净利润现金含量 | X_{16} |

本案例中的变量选择是因变量为上市公司的状态,用 Y 表示;自变量为上表中的财务指标,用 X_i 来表示。

1. 数据来源及说明

案例中的数据分别选取的是沪深两市中的 ST 公司和非 ST 公司,其中包括研究样本和待研究的样本。选取原则主要是配对选取,即对每一个 ST 公司按行业类别、时期和资产规模的原则选取一个非 ST 公司。财务指标的选择为流动比率、总资产周转率、资产净利率和总资产增长率,具体数据参见表 13-3。

2. 判别分析的 SPSS 过程

选择菜单“分析 (Analyze) 分类 (Classify) 判别式 (Discriminant)”即可弹出“判别分析”对话框,如图 13-7 所示。选中变量 Type 到“分组变量 (Grouping Variable)”选项栏中,选中变量“流动比率”、“总资产周转率”、“资产净利率”,以及“总资产增长率”到“自变量 (Independents)”选项栏中。选中“一起输入自变量”选项。

然后单击“分组变量 (Grouping Variable)”选项栏的“定义范围 (Define Range)”按钮,弹出如图 13-8 所示对话框,分别在“最小值”和“最大值”选项栏中填入 1 和 2,然后单击“继续 (Continue)”按钮返回主界面。接着单击主界面中的“统计量 (Statistics)”按钮,弹出如图 13-9 所示对话框,选中均值、单变量 ANOVA、博克斯、费希尔选项,以及“组内相关性”和“组内协方差”选项,然后单击“继续 (Continue)”按钮返回主界面。

单击主界面的“分类 (Classify)”按钮,弹出如图 13-10 所示对话框,设置分类统计量,选中“所有组相等 (All Groups Equal)”选项,“组内 (Within-groups)”选项,以及“摘要表 (Summary table)”选项,然后单击“继续 (Continue)”按钮返回主界面。

最后单击“保存 (Save)”按钮设置保存选项,弹出如图 13-11 所示对话框,选中“预测组成员 (Predicted group membership)”和“判别得分 (Discriminant scores)”选项,然后单击“继续 (Continue)”按钮返回主界面。

表 13-3 上市公司财务数据表

| 公 司 类 型 | 公 司 | 类 型 | 流 动 比 率 | 总资产周转率 | 资产净利率 | 总资产增长率 |
|---------|---------|---------|---------|--------|-------|--------|
| ST | ST 成百 | ST 公司 | 0.6 | 0.03 | 0 | -0.01 |
| | ST 黎明 | ST 公司 | 2 | 0.03 | -0.02 | -0.28 |
| | ST 棱光 | ST 公司 | 1.3 | 0.03 | -0.13 | -0.11 |
| | ST 高斯达 | ST 公司 | 1.8 | 0 | 0 | 0.13 |
| | ST 生态 | ST 公司 | 0.9 | 0.26 | 0.06 | 0.19 |
| | ST 康赛 | ST 公司 | 1.1 | 0.01 | -0.03 | 0 |
| | ST 中燕 | ST 公司 | 0.2 | 0 | -0.23 | -0.26 |
| | ST 鞍一工 | ST 公司 | 0.7 | 0.02 | -0.03 | -0.1 |
| | ST 自仪 | ST 公司 | 0.8 | 0.27 | 0 | -0.3 |
| | ST 达声 | ST 公司 | 0.7 | 0.01 | -0.73 | -0.12 |
| | ST 中华 A | ST 公司 | 0.7 | 0.01 | -0.02 | -0.03 |
| | ST 英达 A | ST 公司 | 0.8 | 0.21 | 0 | -0.07 |
| | ST 中桥 A | ST 公司 | 1 | 0.01 | -0.03 | -0.07 |
| | ST 吉发 | ST 公司 | 1.1 | 0.4 | -0.11 | -0.08 |
| | ST 猴王 | ST 公司 | 0.4 | 0.04 | -0.07 | -0.47 |
| | ST 金马 | ST 公司 | 0.5 | 0.12 | -0.09 | 0.08 |
| | ST 海洋 | ST 公司 | 0 | 0 | 0 | 0.15 |
| | ST 银山 | ST 公司 | 0.5 | 0.21 | -0.02 | -0.09 |
| | ST 合成 | ST 公司 | 1.3 | 0.12 | -0.02 | -0.06 |
| 非 ST | 贵华旅业 | 非 ST 公司 | 0.7 | 0.13 | -0.14 | -0.21 |
| | 江苏吴中 | 非 ST 公司 | 2.1 | 0.49 | 0.06 | 0.17 |
| | 浙江东日 | 非 ST 公司 | 1 | 0.1 | 0.02 | 0.15 |
| | 国际大厦 | 非 ST 公司 | 1.4 | 0.12 | -0.02 | 0.1 |
| | 农产品 | 非 ST 公司 | 0.7 | 0.3 | 0.02 | 0.62 |
| | 浙江富润 | 非 ST 公司 | 1.2 | 0.46 | 0.04 | 0.16 |
| | 上海三毛 | 非 ST 公司 | 1.7 | 0.26 | 0.01 | 0 |
| | 飞彩股份 | 非 ST 公司 | 1.5 | 0.29 | 0.02 | 0.22 |
| | 吴中仪表 | 非 ST 公司 | 2.8 | 0.11 | 0.02 | 0.64 |
| | 夏新电子 | 非 ST 公司 | 1.4 | 0.37 | -0.02 | 0.06 |
| | 济南轻骑 | 非 ST 公司 | 2 | 0.07 | -0.06 | 0 |
| | 北大股份 | 非 ST 公司 | 1.8 | 0.24 | 0.03 | 0.24 |
| | 深宝恒 | 非 ST 公司 | 1.3 | 0.07 | 0.22 | -0.07 |
| | 光彩建设 | 非 ST 公司 | 2.8 | 0.07 | 0 | 0.7 |
| | 大西洋 | 非 ST 公司 | 3.4 | 0.5 | 0.03 | 0 |
| | 西藏圣地 | 非 ST 公司 | 0.6 | 0.07 | 0 | 0.02 |
| | 洞庭水殖 | 非 ST 公司 | 3.8 | 0.11 | 0.01 | 0.06 |
| | 兴发集团 | 非 ST 公司 | 1.3 | 0.24 | 0.02 | -0.03 |
| | 东风药业 | 非 ST 公司 | 4.7 | 0.25 | 0.04 | -0.07 |



图 13-7 “判别设置”对话框



图 13-8 “定义范围”对话框



图 13-9 “统计量设置”对话框



图 13-10 “分类设置”对话框

3. 结果分析

设置完成以后，单击主界面中的“确定”按钮进行判别分析。首先是分析处理统计信息汇总，如图 13-12 所示，给出了观测量个数。



图 13-11 “保存 (Save) 设置”对话框

| 分析个案处理摘要 | | |
|-----------------------------|-----|-------|
| 未加权个案数 | 个案数 | 百分比 |
| 有效 | 38 | 100.0 |
| 排除 | | |
| 缺失或超出范围组代码 | 0 | .0 |
| 至少一个缺失判别变量 | 0 | .0 |
| 既包括缺失或超出范围组代码，也包括至少一个缺失判别变量 | 0 | .0 |
| 总计 | 0 | .0 |
| 总计 | 38 | 100.0 |

图 13-12 汇总信息

然后是组别的统计量。如图 13-13 所示，包括均值、标准差等信息。

图 13-14 是组内的协方差矩阵，图 13-15 是组内均值的显著性检验结果。可以看出，各变量的显著性值均小于 0.05，所以，在显著性水平 0.05 上是比较显著的。

| 组统计 | | | | | |
|------|--------|--------|---------|------------|--------|
| type | | 平均值 | 标准差 | 有效个案数 (成列) | |
| | | | | 未加权 | 加权 |
| 1 | 流动比率 | .8632 | .50023 | 19 | 19.000 |
| | 总资产周转率 | .0937 | .11922 | 19 | 19.000 |
| | 资产净利率 | -.0774 | .17029 | 19 | 19.000 |
| | 总资产增长率 | -.0789 | .16428 | 19 | 19.000 |
| 2 | 流动比率 | 1.9053 | 1.12224 | 19 | 19.000 |
| | 总资产周转率 | .2237 | .14766 | 19 | 19.000 |
| | 资产净利率 | .0158 | .06602 | 19 | 19.000 |
| | 总资产增长率 | .1453 | .25149 | 19 | 19.000 |
| 总计 | 流动比率 | 1.3842 | 1.00661 | 38 | 38.000 |
| | 总资产周转率 | .1587 | .14785 | 38 | 38.000 |
| | 资产净利率 | -.0308 | .13585 | 38 | 38.000 |
| | 总资产增长率 | .0332 | .23834 | 38 | 38.000 |

图 13-13 组别的统计量

| 汇聚组内矩阵 ^a | | | | | |
|---------------------|--------|-------|--------|-------|--------|
| | | 流动比率 | 总资产周转率 | 资产净利率 | 总资产增长率 |
| 协方差 | 流动比率 | .755 | .008 | .011 | .008 |
| | 总资产周转率 | .008 | .018 | .003 | -.001 |
| | 资产净利率 | .011 | .003 | .017 | .004 |
| | 总资产增长率 | .008 | -.001 | .004 | .045 |
| 相关性 | 流动比率 | 1.000 | .068 | .096 | .043 |
| | 总资产周转率 | .068 | 1.000 | .155 | -.022 |
| | 资产净利率 | .096 | .155 | 1.000 | .128 |
| | 总资产增长率 | .043 | -.022 | .128 | 1.000 |

a. 协方差矩阵的自由度为 36。

图 13-14 组内的协方差矩阵

| 组平均值的同等检验 | | | | | |
|-----------|---------------|--------|-------|-------|------|
| | 威尔克 Lambda | F | 自由度 1 | 自由度 2 | 显著性 |
| 流动比率 | .725 | 13.668 | 1 | 36 | .001 |
| 总资产周转率 | .802 | 8.916 | 1 | 36 | .005 |
| 资产净利率 | .879 | 4.943 | 1 | 36 | .033 |
| 总资产增长率 | .773 | 10.585 | 1 | 36 | .002 |

图 13-15 显著性检验

图 13-16 输出的是博克斯检验结果，从结果显著性的值为 0.004 小于 0.05 可以看出，拒绝原假设，从而否定了协方差矩阵相等的假设，下面只能利用分组的协方差矩阵分析 (Separate-groups)。所以，在图 13-10 的选项框中选中分组的协方差矩阵分析选项栏来进行分析，则博克斯 M 检验结果如图 13-17 所示。

下面输出的是判别函数的系数，如图 13-18 所示，从图中可以得到两组类别的 Fisher 判别函数，利用 Fisher 判别函数可以直接计算每个观测属于各组的得分，并把此观测归为得分最高的组别之中，

最后输出的是利用上述所得的 Fisher 判别函数来进行回判的结果，如图 13-19 所示，从图中可以得到回判正确率等信息。对于第一组的观测，总共 19 个观测有 3 个被判为第二

组非 ST 公司, 所以, 回判正确率为 84.2%, 同样第二组的回判正确率为 73.7%。

| 检验结果 | | |
|---------------------|-------|----------|
| 博克斯 M | | 29.643 |
| F | 近似 | 2.605 |
| | 自由度 1 | 10 |
| | 自由度 2 | 6196.016 |
| | 显著性 | .004 |
| 对等同群体协方差矩阵的原假设进行检验。 | | |

图 13-16 博克斯 M 检验结果

| 检验结果 | | |
|----------------------------|-------|----------|
| 博克斯 M | | 2.270 |
| F | 近似 | 2.209 |
| | 自由度 1 | 1 |
| | 自由度 2 | 3888.000 |
| | 显著性 | .137 |
| 对典则判别函数的等同群体协方差矩阵的原假设进行检验。 | | |

图 13-17 调整方法后的博克斯 M 检验结果

| 分类函数系数 | | |
|-----------|--------|--------|
| | type | |
| | 1 | 2 |
| 流动比率 | 1.186 | 2.410 |
| 总资产周转率 | 5.522 | 11.947 |
| 资产净利率 | -5.995 | -3.210 |
| 总资产增长率 | -1.414 | 3.213 |
| (常量) | -1.751 | -4.534 |
| 费希尔线性判别函数 | | |

图 13-18 判别函数的系数

| 分类结果 ^a | | | | | |
|------------------------------|----|---------|------|-------|-------|
| | | 预测组成员信息 | | | 总计 |
| | | type | 1 | 2 | |
| 原始 | 计数 | 1 | 16 | 3 | 19 |
| | 2 | 5 | 14 | 19 | |
| | % | 1 | 84.2 | 15.8 | 100.0 |
| | 2 | 26.3 | 73.7 | 100.0 | |
| a. 正确地对 78.9% 个原始已分组个案进行了分类。 | | | | | |

图 13-19 判别结果

13.3 逐步判别分析

13.3.1 逐步判别的参数设置

如图 13-1 所示, 如果选择“步进法 (Use stepwise method)”选项, 则图 13-1 右上角的“方法 (Method)”按钮会被激活。单击此按钮则会弹出进行逐步判别分析的参数设置对话框, 如图 13-20 所示。各选项栏的功能如下。

(1) 方法 (Method) 栏

此栏用于指定逐步判别分析的方法。

- 威尔克: 每步都选择使总体的 Wilks' lambda 统计量达到最小的变量进入判别分析函数。
- 未解释方差 (Unexplained Variance): 每步都选择使各类别间不可解释的方差和达到最小的变量进入判别函数。
- 马氏距离 (Mahalanobis' Distance): 每步都选择使靠的最近的两个类别的 Mahalanobis 距离达到最大的变量进入判别分析函数。
- 最小 F 比 (Smallest F Value): 选择基于使类间 Mahalanobis 距离计算的一个 F 比率达到最大的变量进入判别函数。
- 拉奥: 选择使统计量 Rao's V 产生最大增量的变量进入判别函数。

(2) 条件 (Criteria) 栏

此栏用于设置逐步判别过程中保留或删除变量的准则。

- 使用 F 值 (Use F value): 使用 F 值, 是系统默认的方法。当变量的 F 值大于进入 (Entry) 值时, 此变量进入模型, 系统默认的进入值为 3.84; 当变量的 F 值小于指定的除去 (Removal) 值时, 该变量从模型中去除, 默认的除去值为 2.71。
- 使用 F 的概率 (Use probability of F): 使用 F 检验的概率值。默认值是 0.05, 默认的删除 (Removal) 值是 0.10。

(3) 显示 (Display) 栏

此栏设置一些输出内容。

- 步骤摘要 (Summary of Steps): 输出逐步判别分析过程中每一步的变量统计信息。
- 成对距离的 F 值 (F for Pairwise Distances): 输出两两类别之间的 F 比率矩阵。



图 13-20 “逐步判别分析参数设置”对话框

13.3.2 实例分析



结果文件——附带光盘“PROGRAM\CH13\实例 13-2”文件夹



动画演示——附带光盘“AVI\实例 13-2.avi”文件

农业是文明和发展的基础, 作为农业人口占大多数的国家, 三农问题解决的好坏, 直接关系到国民经济的持续、稳定、健康发展。随着改革开放的深入发展, 农业也得到了很大的发展, 但是进入 21 世纪以来, 城乡差距的不断扩大, 三农问题也随着农业的发展进入了一个新的阶段。因此, 对农业发展状况的分析研究对经济发展是很重要的。此案例着重对全国各省市农民家庭收支情况进行了一些研究。

数据来源于中国国家统计局。数据主要包括地区、食品、衣着、燃料、住房、生活用品、文化生活, 各个省市地区的农民收支情况的数据集。

研究全国各个省市地区的农民家庭收支的分布规律, 根据国家统计局的调查数据资

料,对 25 个省市的样品进行分析,分成了三个类别,分别是第 1,2,3 组。试对上述数据进行 Fisher 判别分析归类,以求得分类的判别函数。表 13-4 是所要分析的数据集。

表 13-4 全国部分省市地区农民收支情况指标表

| 类 别 | 序 号 | 地 区 | 食 品 | 衣 着 | 燃 料 | 住 房 | 生 活 用 品 | 文 化 生 活 |
|-------|-----|-----|--------|-------|-------|--------|---------|---------|
| 第 1 组 | 1 | 天津 | 135.2 | 36.4 | 10.47 | 44.16 | 36.40 | 3.94 |
| | 2 | 辽宁 | 145.68 | 32.83 | 17.79 | 27.29 | 39.09 | 3.47 |
| | 3 | 吉林 | 159.37 | 33.38 | 18.37 | 11.81 | 25.29 | 5.22 |
| | 4 | 江苏 | 144.98 | 29.12 | 11.67 | 42.6 | 27.30 | 5.74 |
| | 5 | 浙江 | 169.92 | 32.75 | 12.72 | 47.12 | 34.35 | 5.00 |
| | 6 | 山东 | 115.84 | 30.76 | 12.2 | 33.61 | 33.77 | 3.85 |
| 第 2 组 | 7 | 黑龙江 | 116.22 | 29.57 | 13.24 | 13.76 | 21.75 | 6.04 |
| | 8 | 安徽 | 153.11 | 23.09 | 15.62 | 23.54 | 18.18 | 6.39 |
| | 9 | 福建 | 144.92 | 21.26 | 16.96 | 19.52 | 21.75 | 6.73 |
| | 10 | 江西 | 140.54 | 21.59 | 17.64 | 19.19 | 15.97 | 4.94 |
| | 11 | 湖北 | 140.64 | 28.26 | 12.35 | 18.53 | 20.95 | 6.23 |
| | 12 | 湖南 | 164.02 | 24.74 | 13.63 | 22.2 | 18.06 | 6.04 |
| | 13 | 广西 | 139.08 | 18.47 | 14.68 | 13.41 | 20.66 | 3.85 |
| | 14 | 四川 | 137.80 | 20.74 | 11.07 | 17.74 | 16.49 | 4.39 |
| | 15 | 贵州 | 121.67 | 21.53 | 12.58 | 14.49 | 12.18 | 4.57 |
| | 16 | 新疆 | 123.24 | 38.00 | 13.72 | 4.64 | 17.77 | 5.75 |
| | 17 | 河北 | 95.21 | 22.83 | 9.30 | 22.44 | 22.81 | 2.80 |
| 第 3 组 | 18 | 山西 | 104.78 | 25.11 | 6.46 | 9.89 | 18.17 | 3.25 |
| | 19 | 内蒙古 | 128.41 | 27.63 | 8.94 | 12.58 | 23.99 | 3.27 |
| | 20 | 河南 | 101.18 | 23.26 | 8.46 | 20.2 | 20.50 | 4.30 |
| | 21 | 云南 | 124.27 | 19.81 | 8.89 | 14.22 | 15.53 | 3.03 |
| | 22 | 陕西 | 106.02 | 20.56 | 10.94 | 10.11 | 18.00 | 3.29 |
| | 23 | 甘肃 | 95.65 | 16.82 | 5.70 | 6.03 | 12.36 | 4.49 |
| | 24 | 青海 | 107.12 | 16.45 | 8.98 | 5.40 | 8.78 | 5.93 |
| | 25 | 宁夏 | 113.74 | 24.11 | 6.46 | 9.61 | 22.92 | 2.53 |
| 待判组 | 26 | 北京 | 190.33 | 43.77 | 9.73 | 60.54 | 49.01 | 9.04 |
| | 27 | 上海 | 221.11 | 38.64 | 12.53 | 115.65 | 50.82 | 5.89 |
| | 28 | 广东 | 182.55 | 20.52 | 18.32 | 42.4 | 36.97 | 11.68 |

1. 参数设置

首先打开数据集 CH1302,然后进行逐步判别分析设置。选择菜单“分析(Analyze) 分类(Classify) 判别分析(Discriminant)”即可弹出“判别分析”对话框,如图 13-21 所示。选中变量“类别”到“分组变量(Grouping Variable)”选项栏中,选中变量“食品”、“衣着”、“燃料”、“住房”、“生活用品”,以及“文化生活”到“自变量(Independents)”选项栏中。选中“使用步进式方法(Use Stepwise Method)”选项进行逐步分析。

然后单击“分组变量(Grouping Variable)”选项栏的“定义范围(Define Range)”按钮,弹出如图 13-22 所示对话框,分别在“最大”和“最小”选项栏中填入 1 和 3,然后单

击“继续 (Continue)”按钮返回主界面。

然后单击主界面中的“方法 (Method)”按钮，弹出如图 13-23 所示对话框，选择“威尔克”选项，变量进入模型的方法选择“使用 F 的值 (Use value of F)”选项，最后还要选择“步骤摘要 (Summary of steps)”选项给出逐步分析的详细信息。

单击“分类 (Classify)”，图中选择“合并组”、“分组”和“领域图”；显示中选择“摘要表”。

其他选择框中的设置与一般判别分析过程一致，在此不再赘述。

2. 结果分析

设置完成以后，单击主界面中的“确定”按钮进行逐步判别分析，结果如下。首先是基本分析案例的处理结果，如图 13-24 所示。



图 13-21 “逐步分析设置”对话框



图 13-22 “定义范围”对话框



图 13-23 “方法设置”对话框

| 未加权个案数 | 个案数 | 百分比 |
|-----------------------------|-----|-------|
| 有效 | 28 | 100.0 |
| 排除 | | |
| 缺失或超出范围组代码 | 0 | .0 |
| 至少一个缺失判别变量 | 0 | .0 |
| 既包括缺失或超出范围组代码，也包括至少一个缺失判别变量 | 0 | .0 |
| 总计 | 0 | .0 |
| 总计 | 28 | 100.0 |

图 13-24 个案处理结果

然后是各个组别的统计量，包括均值方差等统计信息，如图 13-25 所示。

接着是输出和删除的变量输出结果如图 13-26 所示。图 13-27 是特征值输出和威尔克 Lambda 检验结果，表格给出了第一个判别函数解释了所有变异的 70%，第二个判别函数解释了 30%，威尔克 Lambda 检验结果用于检验各个判别函数有无统计学上的显著意义，由于其对应的显著性值均小于 0.05，所以说这两个判别函数作用都是显著成立的。

然后是标准化判别函数的系数输出，如图 13-28 所示。图 13-29 是结构矩阵的输出结果。结构矩阵给出了判别变量和标准化判别函数之间的相关性数据，可以用来判断各个判别函数受哪些判别变量的影响较大。

| 组统计 | | | | | |
|-----|------|----------|----------|------------|--------|
| 类别 | | 平均值 | 标准差 | 有效个案数 (成列) | |
| | | | | 未加权 | 加权 |
| 1 | 食品 | 162.7756 | 31.99242 | 9 | 9.000 |
| | 衣着 | 33.1300 | 6.47368 | 9 | 9.000 |
| | 燃料 | 13.7556 | 3.44143 | 9 | 9.000 |
| | 住房 | 47.2422 | 29.04810 | 9 | 9.000 |
| | 生活用品 | 37.0000 | 8.57781 | 9 | 9.000 |
| | 文化生活 | 5.9811 | 2.70314 | 9 | 9.000 |
| 2 | 食品 | 134.2227 | 18.97577 | 11 | 11.000 |
| | 衣着 | 24.5527 | 5.52432 | 11 | 11.000 |
| | 燃料 | 13.7082 | 2.45699 | 11 | 11.000 |
| | 住房 | 17.2236 | 5.43457 | 11 | 11.000 |
| | 生活用品 | 18.7791 | 3.17372 | 11 | 11.000 |
| | 文化生活 | 5.2482 | 1.23325 | 11 | 11.000 |
| 3 | 食品 | 110.1463 | 11.28511 | 8 | 8.000 |
| | 衣着 | 21.7188 | 3.98713 | 8 | 8.000 |
| | 燃料 | 8.1038 | 1.75004 | 8 | 8.000 |
| | 住房 | 11.0050 | 4.74374 | 8 | 8.000 |
| | 生活用品 | 17.5313 | 5.17586 | 8 | 8.000 |
| | 文化生活 | 3.7613 | 1.08892 | 8 | 8.000 |
| 总计 | 食品 | 136.5214 | 30.12765 | 28 | 28.000 |
| | 衣着 | 26.5000 | 7.12829 | 28 | 28.000 |
| | 燃料 | 12.1221 | 3.63836 | 28 | 28.000 |
| | 住房 | 25.0957 | 22.67930 | 28 | 28.000 |
| | 生活用品 | 24.2793 | 10.59401 | 28 | 28.000 |
| | 文化生活 | 5.0589 | 1.95775 | 28 | 28.000 |

图 13-25 各个组别的统计信息

| 威尔克 Lambda | | | | | | | | | |
|------------|-----|--------|-------|-------|-------|--------|-------|--------|------|
| 步骤 | 变量数 | Lambda | 自由度 1 | 自由度 2 | 自由度 3 | 统计 | 精确 F | | |
| | | | | | | | 自由度 1 | 自由度 2 | 显著性 |
| 1 | 1 | .289 | 1 | 2 | 25 | 30.697 | 2 | 25.000 | .000 |
| 2 | 2 | .142 | 2 | 2 | 25 | 19.862 | 4 | 48.000 | .000 |

图 13-26 样本协方差矩阵相等的检验结果

然后是费希尔判别函数的系数,如图 13-30 所示。费希尔判别函数可以直接计算各个观测量的得分,并根据得分归类。

图 13-31 是逐步判别分析的边界图的输出,边界图是根据典型判别函数,按照观测量和各类别重心的距离在平面上划分类别区域,某观测变量按照判别函数计算的坐标落在哪个区域,则就会被归为那一类别之中。

| 特征值 | | | | |
|-----|--------------------|-------|-------|-------|
| 函数 | 特征值 | 方差百分比 | 累计百分比 | 典型相关性 |
| 1 | 3.464 ^a | 85.7 | 85.7 | .881 |
| 2 | .579 ^a | 14.3 | 100.0 | .606 |

a. 在分析中使用了前 2 个典则判别函数。

| 威尔克 Lambda | | | | |
|------------|------------|--------|-----|------|
| 函数检验 | 威尔克 Lambda | 卡方 | 自由度 | 显著性 |
| 1 直至 2 | .142 | 47.848 | 4 | .000 |
| 2 | .633 | 11.194 | 1 | .001 |

图 13-27 特征值输出和威尔克 Lambda 检验结果

| 标准化典则判别函数系数 | | |
|-------------|------|-------|
| | 函数 | |
| | 1 | 2 |
| 燃料 | .607 | .828 |
| 生活用品 | .944 | -.403 |

图 13-28 标准化判别函数的系数

| 结构矩阵 | | |
|-------------------|-------------------|--------------------|
| | 函数 | |
| | 1 | 2 |
| 生活用品 | .806 ^a | -.591 |
| 食品 ^b | .567 ^a | .003 |
| 文化生活 ^b | .231 ^a | .164 |
| 燃料 | .393 | .920 ^a |
| 住房 ^b | .452 | -.526 ^a |
| 衣着 ^b | .255 | -.480 ^a |

判别变量与标准化典则判别函数之间的汇聚组内相关性
变量按函数内相关性的绝对大小排序。

*. 每个变量与任何判别函数之间的最大绝对相关性

b. 在分析中未使用此变量。

图 13-29 结构矩阵

| 分类函数系数 | | | |
|--------|---------|---------|---------|
| | 类别 | | |
| | 1 | 2 | 3 |
| 燃料 | 2.616 | 2.332 | 1.476 |
| 生活用品 | 1.321 | .773 | .650 |
| (常量) | -43.531 | -24.337 | -12.778 |

费希尔线性判别函数

图 13-30 费希尔判别函数的系数

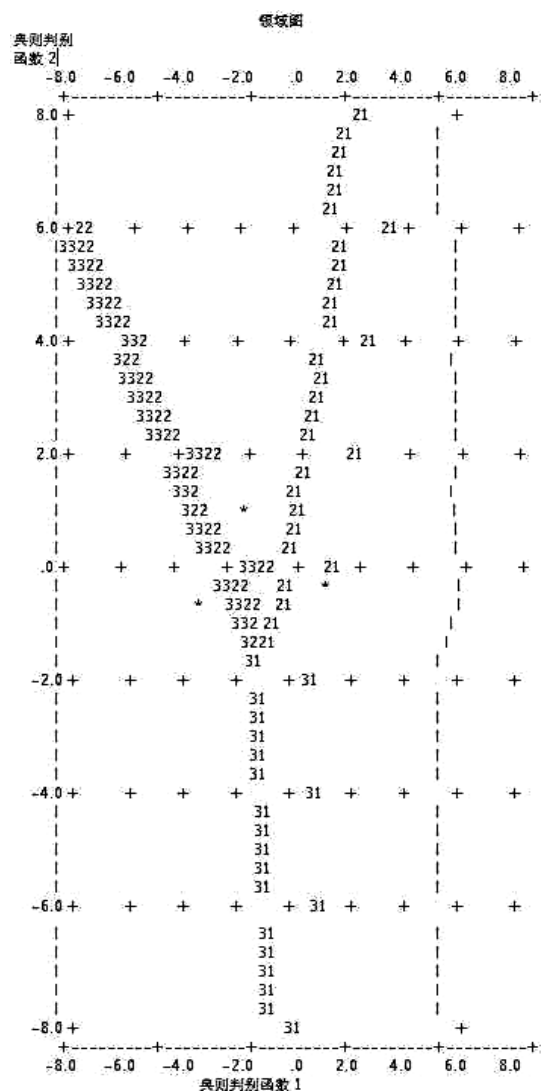


图 13-31 边界图

然后是三个类别的散点图形,如图 13-32 所示,从图中可以直观地看出各个观测的落点。

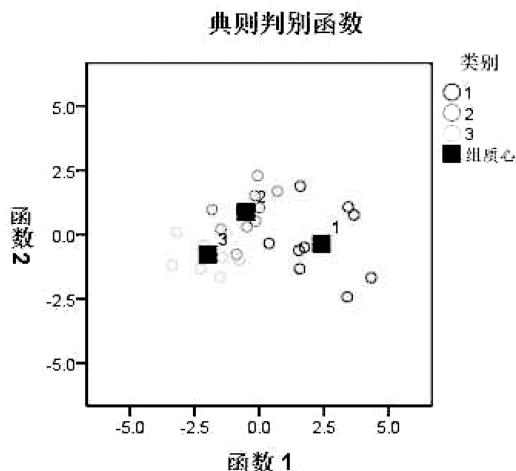


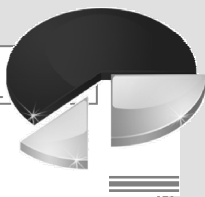
图 13-32 散点图形

最后输出的是回判输出表,如图 13-33 所示,输出的是利用上述所得的 Fisher 判别函数来进行回判的结果,从图中可以得到回判正确率等信息。对于第 1 组的观测,总共 9 个观测,有 2 个被判为第 2 组,所以回判正确率为 77.8%,同样第 2 组的回判正确率为 81.8%,第 3 组的回判正确率为 87.5%。

| 分类结果 ^a | | | | | | |
|-------------------|----|---------|------|------|------|-------|
| | | 预测组成员信息 | | | | 总计 |
| | | 类别 | 1 | 2 | 3 | |
| 原始 | 计数 | 1 | 7 | 2 | 0 | 9 |
| | | 2 | 0 | 9 | 2 | 11 |
| | | 3 | 0 | 1 | 7 | 8 |
| | % | 1 | 77.8 | 22.2 | .0 | 100.0 |
| | | 2 | .0 | 81.8 | 18.2 | 100.0 |
| | | 3 | .0 | 12.5 | 87.5 | 100.0 |

a. 正确地 对 82.1% 个原始已分组个案进行了分类。

图 13-33 判别结果



第 14 章 因子分析

在实际问题中，研究多指标问题是经常遇到的，然而在多数情况下，不同指标之间是有一定相关性的。由于指标较多，再加上指标之间有一定的相关性，势必增加了分析问题的复杂性。而在商业经济中，用主成分分析可将复杂的一些数据综合成几个商业指数形式，如生活费用指数，物价指数，商业活动指数等。如上所述，现实中往往希望综合使用这些指标。这时，因子分析方法可以把数据的维数降低，同时又尽量不损失数据中的信息。

因子分析是主成分分析的推广，它也是从研究相关矩阵内部的依赖关系出发，把一些具有错综复杂关系的变量归结为少数几个综合变量的一种多变量统计分析方法。本章将利用 SPSS 软件系统进行因子分析。



本讲内容

- 因子分析
- SPSS 软件进行因子分析

14.1 因子分析简介

主成分分析通过线性组合将原变量综合成几个主成分，用较少的综合指标来代替原来较多的指标（变量）。在多变量分析中，某些变量间往往存在相关性。是什么原因使变量间有关联呢？是否存在不能直接观测到的但影响可观测变量变化的公共因子？因子分析法（Factor Analysis）就是寻找这些公共因子的模型分析方法，它是在主成分的基础上构筑若干意义较为明确的公因子，以它们为框架分解原变量，以此考察原变量间的联系与区别。

例如，随着年龄的增长，儿童的身高、体重会随着变化，具有一定的相关性，身高和体重之间为何会有相关性呢？因为存在着一个同时支配或影响着身高与体重的生长因子。那么，能否通过对多个变量的相关系数矩阵的研究，找出同时影响或支配所有变量的共性因子呢？因子分析就是从大量的数据中“由表及里”、“去粗取精”，寻找影响或支配变量的多变量统计方法。

可以说，因子分析是主成分分析的推广，也是一种把多个变量化为少数几个综合变量的多变量分析方法，其目的是用有限个不可观测的隐变量来解释原始变量之间的相关关系。

因子分析主要用于：减少分析变量个数；通过对变量间相关关系探测，将原始变量进行分类。即将相关性高的变量分为一组，用共性因子代替该组变量。

14.1.1 因子分析的基本原理

因子分析法是从研究变量内部相关的依赖关系出发，把一些具有错综复杂关系的变量归结为少数几个综合因子的一种多变量统计分析方法。它的基本思想是将观测变量进行分类，将相关性较高，即联系比较紧密的分在同一类中，而不同类变量之间的相关性则较低，那么每一类变量实际上就代表了一个基本结构，即公共因子。对于所研究的问题就是试图用最少数个数的不可测的公共因子的线性函数与特殊因子之和来描述原来观测的每一分量。

1. 因子分析模型

首先，定义观测值所构成的矩阵如下：

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

式中： n 为样本观测的次数， p 为变量数， $X_i = (x_{1i}, x_{2i}, \cdots, x_{ni})'$ ， $i = 1, 2, 3, \cdots, p$ 。然后将 X 中的数据进行标准化处理，则处理后的变量的方差为 1，均值为 0。为了叙述简单，假设经过标准化后的矩阵仍记为 X 。所以相关系数矩阵为

$$R = X'X$$

设 R 的 p 个非负特征值为 $\lambda_1, \lambda_2, \cdots, \lambda_p$ 。记对应于特征值的正交特征矢量矩阵如下：

$$U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

令 $F = UX'$ ，则有下面等式：

$$FF' = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

上式中 F 为主因子阵，并且 $F_i = U_i X'$ ， $i = 1, 2, 3, \cdots, p$ ，即 F_i 为第 i 个样品的主因子得分。下面选择 $m(m < p)$ 个主因子，根据变量的相关阵选出第一主因子 F_1 ，使其在各个变量的公共因子方差中所占的方差贡献最大，然后就消去此因子的影响；再从剩余的相关阵中选出与 F_1 不相关的因子 F_2 ，以次类推，直到各个变量公共因子方差被分解完毕为止。

由 $m(m < p)$ 个主因子将 U 矩阵分为两部分：

$$U = [U_1, U_2, \cdots, U_m, U_{m+1}, \cdots, U_p] = [U_{(1)}, U_{(2)}]$$

式中： $U_{(1)}$ 为 $p \times m$ 矩阵， $U_{(2)}$ 为 $p \times (p - m)$ 矩阵。

由 $F = UX'$ ，可知 $X = U'F$ ，再令 $F = [F_{(1)}, F_{(2)}]$ ，其中 $F_{(1)}$ 为 $m \times n$ 矩阵， $F_{(2)}$ 为 $(p-m) \times n$ 矩阵。则有

$$X = U'_{(1)} F_{(1)} + U'_{(2)} F_{(2)}$$

$U'_{(1)} F_{(1)}$ 为 m 个主因子所能解释的部分， $U'_{(2)} F_{(2)}$ 为其残差部分。记残差为 ε 。所以，就可以得到因子模型如下：

$$X = U'_{(1)} F_{(1)} + \varepsilon$$

式中， $U_{(1)}$ 为因子负荷矩阵， $F_{(1)}$ 为主因子， ε 为特殊因子。

如果略去特殊因子，则因子模型就为

$$\begin{cases} X_1 = u_{11}F_1 + u_{12}F_2 + \cdots + u_{1m}F_m \\ X_2 = u_{21}F_1 + u_{22}F_2 + \cdots + u_{2m}F_m \\ \vdots \\ X_p = u_{p1}F_1 + u_{p2}F_2 + \cdots + u_{pm}F_m \end{cases}$$

令 $a_{ij} = u_{ij}\lambda_j^{1/2}$ ，所以因子负荷矩阵为 $A = (a_{ij})_{p \times m}$ ，因此，因子分析的数学模型为

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\ \vdots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p \end{cases}$$

式中， $F = (X_1, X_2, \dots, X_p)'$ ，称 X 为公共因子， $A = (a_{ij})_{p \times m}$ 为因子载荷矩阵， a_{ij} 为因子负荷，它是第 i 个变量在第 j 个公共因子上的负荷，反映了第 i 个变量在第 j 个变量上的相对重要性。残差 ε 为特殊因子，相互独立，且服从正态分布 $N(0, \sigma_i^2)$ 。

2. 因子旋转

得出因子模型以后，下面就会对公共因子进行解释，为了更好地解释公共因子，减少解释的主观性，主要采用因子旋转方法，得到比较满意的主因子。从线性代数的角度看，因子旋转即是非奇异的线性变换。

上述的因子模型写成矩阵形式如下：

$$X = AF + \varepsilon$$

进行因子旋转的目的就是要使因子负荷矩阵中因子负荷的平方值向 0 和 1 两个方向分化，因子旋转的方法主要有正交旋转和斜交旋转方法。这里要注意的是，采用正交旋转得到的因子也是不相关的，但斜交旋转得到的因子是相关的。SPSS 中可供选择的因子旋转方法主要有方差最大正交旋转方法、正交旋转法、平衡法等。

3. 计算因子得分

因子分析是将变量表示为公共因子的线性组合，如果将因子表示为变量的线性组合，即

$$\begin{cases} f_1 = \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p \\ f_2 = \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p \\ \vdots \\ f_m = \beta_{m1}x_1 + \beta_{m2}x_2 + \cdots + \beta_{mp}x_p \end{cases}$$

即为因子得分函数,利用得分函数可以计算每个样本的因子得分。由于上式的个数少于变量个数,因此,只能在最小二乘的意义下对因子得分进行估计。估计因子得分的方法较多,常用的有回归估计法、Bartlett 估计法和 Thomson 估计法。

14.1.2 因子分析的基本步骤和过程

因子分析的核心问题有两个:一是如何构造因子变量;二是如何对因子变量进行命名解释。因此,因子分析的基本步骤和解决思路就是围绕这两个核心问题展开的。因子分析常常有以下四个基本步骤:

确认待分析的原变量是否适合作因子分析。

构造因子变量。

利用旋转方法使因子变量更具有可解释性。

计算因子变量得分。

综合上节中的因子分析的数学模型,可以知道因子分析的基本计算过程如下:

将原始数据标准化,以消除变量间在数量级和量纲上的不同。

求标准化数据的相关矩阵。

求相关矩阵的特征值和特征矢量。

计算方差贡献率与累积方差贡献率。

确定因子,设 F_1, F_2, \dots, F_p 为 p 个因子,其中,当前 m 个因子包含的数据信息总量不低于 80% 时,可以取前 m 个因子来反映原评价指标。

因子旋转,若所得的 m 个因子无法确定或其实际意义不是很明显,这时需将因子进行旋转以获得较为明显的实际含义。

用原始指标的线性组合来求得各个因子的得分,采用回归估计法、Bartlett 估计法或 Thomson 估计法计算因子得分。

综合得分,以各因子的方差贡献率为权,由各因子的线性组合得到综合评价指标函数,即

$$F = \frac{w_1 F_1 + w_2 F_2 + \dots + w_m F_m}{w_1 + w_2 + \dots + w_m}$$

式中, w_i 为旋转前或者旋转后因子的方差贡献率。

得分排序:利用综合得分可以得到得分名次。

14.2 SPSS 因子分析

14.2.1 SPSS 因子分析的参数设置

选择菜单“分析（Analyze） 降维（Data Reduction） 因子分析（Factor）”，则系统执行因子分析过程，弹出如图 14-1 所示的对话框。各选项框的具体功能如下。

1. 变量设置

进行因子分析之前要进行变量选择设置，各选项栏功能如下所述。

- 变量（Variables）：用于从图 14-1 左边的变量框中选入待分析的原始变量。
- 选择变量（Selection Variable）：用于从图 14-1 左边的变量框中选入过滤样本子集的变量。选入变量以后，则激活“值（Value）”按钮，单击弹出如图 14-2 所示对话框。在“选定变量的值（Value for Selection Variable）”输入框中指定变量的某个取值，当变量取这个值时才会进行分析过程。



图 14-1 “因子分析（Factor）”对话框



图 14-2 “值（Value）”对话框

2. 描述（Descriptives）设置

单击图 14-1 右上角的“描述（Descriptives）”按钮，则弹出如图 14-3 所示的描述“统计选项设置”对话框。单击“继续（Continue）”按钮则返回主界面。

（1）统计量（Statistics）栏

此栏选择要输出的统计量。

- 单变量描述性（Univariate Descriptive）：单变量描述统计量，输出变量的均值、标准差和有效取值个数。
- 初始解（Initial Solution）：输出包括初始公共因子、初始特征根和初始方差贡献率等信息。

（2）相关性矩阵（Correlation Matrix）栏

此栏输出有关相关矩阵等信息。

- 系数（Coefficients）：输出初始分析变量之间的相关系数矩阵。

- 逆 (Inverse): 相关系数矩阵的行列式。
- 显著性水平 (Significance Levels): 显著性水平, 输出每个相关系数关于单侧假设检验的显著性 P 值。
- 再生 (Reproduced): 再生相关矩阵, 输出因子分析后的相关矩阵。
- 反映像 (Anti-image): 输出偏相关系数的负数, 反象协方差矩阵包括偏协方差的负数。
- KMO 和巴特利特的球形度检验 (KMO and Bartlett's test of sphericity): 输出 KMO 检验和球形 Bartlett 检验。

3. 提取 (Extraction) 设置

单击图 14-1 右上角的“提取 (Extraction)”按钮, 则弹出如图 14-4 所示的“提取选项设置”对话框。单击“继续 (Continue)”按钮则返回主界面。



图 14-3 “描述统计 (Descriptive) 设置”对话框



图 14-4 “提取 (Extraction) 选项设置”对话框

方法 (Method) 下拉菜单, 设置公共因子的提取方法, 如图 14-5 所示。

- 主成分法 (Principal Components), 该方法假设变量是因子的线性组合。
 - 未加权最小平方方法 (Unweighted Least Square), 该方法使得观测的相关矩阵和再生的相关矩阵之差的平方和最小。
 - 广义最小平方方法 (Generalized Least Square)
 - 最大似然 (K) (Maximun Likelihood)
 - 主轴因式分解 (Principal Axis Factoring)
 - Alpha 因式分解 (Alpha Factoring)
 - 映像因式分解 (Image Factoring)
- 分析 (Analyze) 栏: 此栏用于计算公共因子矩阵。
- 相关性矩阵 (Correlation Matrix): 以分析变量的相关矩阵作为提取公共因子的依据。
 - 协方差矩阵 (Covariance Matrix): 以分析变量的协方差矩阵作为提取公共因子的依据。
- 输出 (Display) 栏: 此栏用于选择与因子提取有关的输出选项。
- 未旋转因子解 (Unrotated Factor Solution): 输出未经旋转的因子载荷矩阵。

- 碎石图 (Scree Plot): 输出以按特征值大小排列的因子序号为横轴、特征值为纵轴所绘制的碎石图。

提取 (Extract) 栏: 此栏用于设置提取公共因子的规则。

- 基于特征值 (Eigenvalues over): 指定需要提取的公共因子的最小特征值, 默认为 1。
- 因子的固定数量 (Number of factors): 指定要提取的公共因子的数目。

最大收敛性迭代次数 (Maximum Iterations for Convergence) 栏: 指定因子提取算法收敛的最大迭代次数, 系统默认为 25。

4. 旋转 (Rotation) 设置

单击图 14-1 右上角的“旋转 (Rotation)”按钮, 则弹出如图 14-6 所示的“旋转选项设置”对话框。单击“继续”按钮则返回主界面。

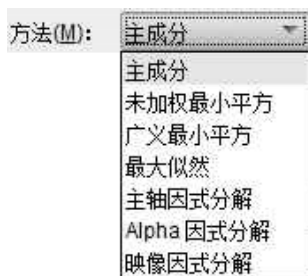


图 14-5 设置公因子的提取方法



图 14-6 “旋转 (Rotation) 选项设置”对话框

方法 (Method) 栏: 此栏用于选择因子旋转的方法。

- 无 (None): 不进行旋转操作。
- 最大方差法 (Varimax): 方差最大旋转方法。
- 直接斜交方法 (Direct Oblimin): 直接斜交旋转。旋转后会激活其下的 Delta 选项框, 用于指定 Delta 参数, 默认为 0。
- 四次幂极大法 (Quartimax): 四次最大正交旋转。
- 等量最大法 (Equamax): 平均正交旋转。
- 最优斜交法: 斜交旋转方法, 旋转后会激活其下的 Kappa 选项框, 默认为 4。

输出 (Display) 栏: 在此选择有关因子旋转的输出。

- 旋转后的解 (Rotated Solution): 选择结果, 当指定某种旋转方法后会被激活。
- 载荷图 (Loading Plot(s)): 因子载荷散点图。

最大收敛性迭代次数 (Maximum Iterations for Convergence): 指定因子旋转收敛的最大迭代次数, 系统默认为 25。

5. 得分 (Scores) 设置

单击图 14-1 右上角的“得分 (Scores)”按钮, 则弹出如图 14-7 所示的“得分选项设置”对话框。单击“继续 (Continue)”按钮则返回主界面。

保存为变量 (Save as Variables) 栏: 此栏表示把每个因子得分作为一个新变量保存到当前数据集中。

方法 (Method) 栏: 此栏用于设置估计因子得分系数的方法。

- 回归 (Regression): 回归方法。
- 巴特利特: Bartlett 方法, 因子得分为 0。
- 安德森-鲁宾: 此方法是 Bartlett 方法的调整, 可以保证估计因子的正交性, 其因子得分的均值为 0, 标准差为 1。

显示因子得分系数矩阵 (Display factor score coefficient matrix): 输出标准化的因子得分系数矩阵, 对原始变量进行标准化以后, 可以根据该矩阵计算各观测量的因子得分。

6. 选项 (Options) 设置

单击图 14-1 右上角的“选项 (Options)”按钮, 则弹出如图 14-8 所示的“选项设置”对话框。单击“继续 (Continue)”按钮则返回主界面。



图 14-7 “得分 (Scores) 设置”对话框



图 14-8 “选项 (Options) 设置”对话框

缺失值 (Missing Values) 栏: 此栏用于设置对缺失值的处理方法。

- 成列排除个案 (Exclude Cases Listwise): 当选入了多个变量进行分析时, 只要其中的某个变量取缺失值, 就在所有分析过程中将对应的记录删除。
- 成对排除个案 (Exclude Cases Pairwise): 成对剔除带有缺失值的观测值, 在计算某个特定的统计量时, 只有当前用到的某个变量取缺失值时, 才将相应的记录删除。
- 替换为平均值 (Replace with Mean): 用变量的均值代替其缺失值。

系数显示格式 (Coefficient Display Format) 栏: 此栏用于选择载荷系数的显示格式。

- 按大小排序 (Sorted by Size): 载荷系数按照取值大小排序。
- 排除小系数 (Suppress Absolute Values Less than): 不显示绝对值小于指定值的载荷系数, 选中此项后, 其后的选项框中的默认值是 0.10, 用户也可以填入 0~1 之间的数作为临界值。

14.2.2 实例分析



结果文件

——附带光盘“PROGRAM\CH14\实例 14-1”文件夹



动画演示

——附带光盘“AVI\实例 14-1.avi”文件

本实例使用 SPSS 中自带的数据集 car_sales.sav，此数据集是关于汽车销售的数据，共包含 26 个变量，如 manufact、model、sales、resale、type、price 等变量，下面就利用因子分析来分析此数据集。数据集如图 14-9 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|----------|-----|----|------|---------------------|----------------|----|----|----|----|----|
| 1 | manufact | 字符串 | 13 | 0 | Manufacturer | 无 | 无 | 7 | 左 | 名义 | 输入 |
| 2 | model | 字符串 | 17 | 0 | Model | 无 | 无 | 10 | 左 | 名义 | 输入 |
| 3 | sales | 数字 | 11 | 3 | Sales in thousa... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 4 | resale | 数字 | 11 | 3 | 4-year resale va... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 5 | type | 数字 | 11 | 0 | Vehicle type | {0, Automob... | 无 | 8 | 右 | 有序 | 输入 |
| 6 | price | 数字 | 11 | 3 | Price in thousa... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 7 | engine_s | 数字 | 11 | 1 | Engine size | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 8 | horsepow | 数字 | 11 | 0 | Horsepower | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 9 | wheelbas | 数字 | 11 | 1 | Wheelbase | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 14-9 数据集 car_sales.sav

1. 参数设置

选择菜单“分析（Analyze）降维（Data Reduction）因子分析（Factor）”，则系统执行因子分析过程，弹出如图 14-10 所示的对话框。选择变量 type、price、engine_s、horsepow、wheelbase、width、length、curb_wgt、fuel_cap、mpg 并选入“变量（Variables）”选项栏中。

然后单击图 14-10 中的“提取（Extraction）”按钮，弹出如图 14-11 所示对话框，此对话框用于设置计算公因子的方法。选中“碎石图 Scree plot”选项框，然后单击“继续”按钮返回主界面。



图 14-10 因子（Factor）主界面



图 14-11 “提取（Extraction）设置”对话框

单击主界面因子分析对话框中的“旋转（Rotation）”按钮，用于设置因子旋转，如

图 14-12 所示。在“方法 (Method)”选项栏中选中“最大方差法 (Varimax)”选项,然后单击“继续”按钮返回主界面。

单击主界面因子分析对话框中的“得分 (Scores)”按钮,用于设置因子得分,如图 14-13 所示。选中“保存为变量 (Save as Variables)”选项和“显示因子得分系数矩阵 (Display Factor Score Coefficient Matrix)”选项,然后单击“继续”按钮返回主界面。



图 14-12 “旋转 (Rotation) 设置”对话框

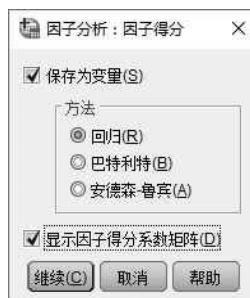


图 14-13 “得分 (Scores) 设置”对话框

2. 结果分析

设置好上述参数以后,返回主界面,单击“确定”按钮进行因子分析,结果在 SPSS 输出窗口中显示。首先是变量共同度表,如图 14-14 所示。表中给出了提取公共因子前后各变量的共同度,它是衡量公因子相对重要性指标,例如,表格第一行数据说明变量 Vehicle type 的共同度为 0.930,即提取的公因子对变量 Vehicle type 的方差贡献率为 93%。

| 公因子方差 | | |
|--------------------|-------|------|
| | 初始 | 提取 |
| Vehicle type | 1.000 | .930 |
| Price in thousands | 1.000 | .876 |
| Engine size | 1.000 | .843 |
| Horsepower | 1.000 | .933 |
| Wheelbase | 1.000 | .881 |
| Width | 1.000 | .776 |
| Length | 1.000 | .919 |
| Curb weight | 1.000 | .891 |
| Fuel capacity | 1.000 | .861 |
| Fuel efficiency | 1.000 | .860 |
| 提取方法: 主成分分析法。 | | |

图 14-14 变量共同度表

然后是主成分结果,如图 14-15 所示,表中列出了所有主成分,且按照特征根的从大到小次序排序,从图中可以看出第一主成分特征根为 5.994,方差贡献率是 59.938%,前三个因子的方差累积贡献率是 87.709%。根据因子提取条件,特征值大于 1,所以选择前三个因子。

| 总方差解释 | | | | | | | | | |
|-------|-------|--------|---------|---------|--------|--------|---------|--------|--------|
| 成分 | 初始特征值 | | | 提取载荷平方和 | | | 旋转载荷平方和 | | |
| | 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 | 累积 % |
| 1 | 5.994 | 59.938 | 59.938 | 5.994 | 59.938 | 59.938 | 3.220 | 32.199 | 32.199 |
| 2 | 1.654 | 16.545 | 76.482 | 1.654 | 16.545 | 76.482 | 3.134 | 31.344 | 63.543 |
| 3 | 1.123 | 11.227 | 87.709 | 1.123 | 11.227 | 87.709 | 2.417 | 24.166 | 87.709 |
| 4 | .339 | 3.389 | 91.098 | | | | | | |
| 5 | .254 | 2.541 | 93.640 | | | | | | |
| 6 | .199 | 1.994 | 95.633 | | | | | | |
| 7 | .155 | 1.547 | 97.181 | | | | | | |
| 8 | .130 | 1.299 | 98.480 | | | | | | |
| 9 | .091 | .905 | 99.385 | | | | | | |
| 10 | .061 | .615 | 100.000 | | | | | | |

提取方法：主成分分析法。

图 14-15 主成分表

图 14-16 是因子分析的碎石图，是按照特征根大小排列的主成分散点图，图中纵坐标为特征值，横坐标是因子数，从图中可以看出除了前三个主成分外，其他的主成分特征根都很小。

图 14-17 给出的是因子载荷矩阵，用来反映各个变量的变异可以主要由哪些因子解释，因子表达式如下。

$$\text{Vehicle type} = 0.471F_1 + 0.533F_2 - 0.651F_3$$

$$\text{Price in thousands} = 0.580F_1 - 0.729F_2 - 0.092F_3$$

$$\text{Engine size} = 0.871F_1 - 0.290F_2 + 0.018F_3$$

$$\text{Horsepower} = 0.740F_1 - 0.618F_2 + 0.058F_3$$

$$\text{Wheelbase} = 0.732F_1 + 0.480F_2 + 0.340F_3$$

$$\text{Width} = 0.821F_1 + 0.114F_2 + 0.298F_3$$

$$\text{Length} = 0.719F_1 + 0.304F_2 + 0.556F_3$$

$$\text{Curb weight} = 0.934F_1 + 0.063F_2 - 0.121F_3$$

$$\text{Fuel capacity} = 0.885F_1 + 0.184F_2 - 0.210F_3$$

$$\text{Fuel efficiency} = -0.863F_1 + 0.004F_2 + 0.339F_3$$

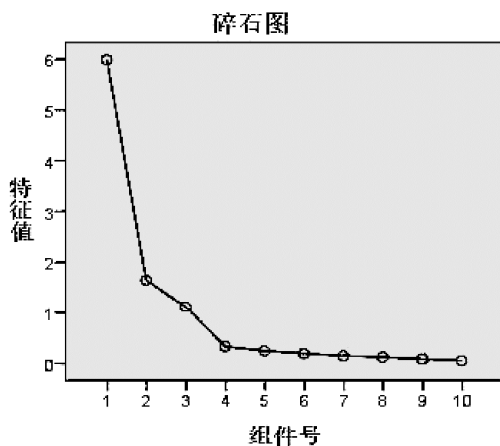


图 14-16 因子分析的碎石图

| 成分矩阵 ^a | | | |
|--------------------|-------|-------|-------|
| | 成分 | | |
| | 1 | 2 | 3 |
| Vehicle type | .471 | .533 | -.651 |
| Price in thousands | .580 | -.729 | -.092 |
| Engine size | .871 | -.290 | .018 |
| Horsepower | .740 | -.618 | .058 |
| Wheelbase | .732 | .480 | .340 |
| Width | .821 | .114 | .298 |
| Length | .719 | .304 | .556 |
| Curb weight | .934 | .063 | -.121 |
| Fuel capacity | .885 | .184 | -.210 |
| Fuel efficiency | -.863 | -.004 | .339 |

提取方法：主成分分析法。
a. 提取了 3 个成分。

图 14-17 因子载荷矩阵

如图 14-18 所示的是因子得分系数矩阵，通过此表就可以得到用各个变量的线性组合表达的主成分，表达式如下。

$$F1 = -0.173X1 + 0.414X2 + 0.226X3 + 0.368X4 - 0.177X5 + 0.011X6 - 0.105X7 + 0.070X8 + 0.012X9 - 0.107X10$$

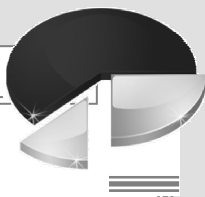
$$F2 = -0.194X1 - 0.179X2 + 0.028X3 - 0.046X4 + 0.397X5 + 0.289X6 + 0.477X7 + 0.043X8 + 0.017X9 + 0.108X10$$

$$F3 = 0.615X1 - 0.081X2 - 0.016X3 - 0.139X4 - 0.042X5 - 0.102X6 - 0.234X7 + 0.175X8 + 0.262X9 - 0.298X10$$

| 成分得分系数矩阵 | | | |
|--------------------|-------|-------|-------|
| | 成分 | | |
| | 1 | 2 | 3 |
| Vehicle type | -.173 | -.194 | .615 |
| Price in thousands | .414 | -.179 | -.081 |
| Engine size | .226 | .028 | -.016 |
| Horsepower | .368 | -.046 | -.139 |
| Wheelbase | -.177 | .397 | -.042 |
| Width | .011 | .289 | -.102 |
| Length | -.105 | .477 | -.234 |
| Curb weight | .070 | .043 | .175 |
| Fuel capacity | .012 | .017 | .262 |
| Fuel efficiency | -.107 | .108 | -.298 |

提取方法：主成分分析法。
旋转方法：凯撒-梅尔-奥克斯-穆恩最大方差法。
组件得分。

图 14-18 因子得分系数矩阵



第 15 章 对应分析

对应分析 (Correspondence Analysis) 是由法国人 Benzenci 于 1970 年提出的一种多元相依变量统计分析技术, 通过分析由定性变量构成的交互汇总表来揭示变量间的联系。更重要的是对应分析法是一种视觉化的数据分析方法, 用户可以直观地观察数据分析的结果。对应分析在很多方面都有应用, 例如, 市场细分、城镇居民消费结构分析、临床医学等方面。

本章将详细叙述对应分析的理论, 并利用 SPSS 过程实现对应分析。



本讲内容

- 对应分析的基本原理
- 简单对应分析
- Optimal Scaling 过程

15.1 对应分析的基本原理

对应分析也称关联分析、R-Q (样本-变量) 型因子分析, 是近年新发展起来的一种多元相依变量统计分析技术, 通过分析由定性变量构成的交互汇总表来揭示变量间的联系。可以揭示同一变量的各个类别之间的差异, 以及不同变量各个类别之间的对应关系。它是一种视觉化的数据分析方法, 它能够将几组看不出任何联系的数据, 通过视觉上可以接受的定位图展现出来。广泛应用在市场细分、产品定位、地质研究, 以及计算机工程等领域中。

对应分析起初在法国和日本最为流行, 然后引入到美国。对应分析法是在 R 型和 Q 型因子分析的基础上发展起来的一种多元统计分析方法, 因此又称 R-Q 型因子分析。在因子分析中, 如果研究的对象是样品, 则需采用 Q 型因子分析; 如果研究对象是变量, 则需采用 R 型因子分析。但是, 这两种分析方法往往是相互对立的, 必须分别对样品和变量进行处理。因此, 因子分析对于分析样品的属性和样品之间的内在联系, 就比较困难, 因为样品的属性是变值, 而样品却是固定的。于是就产生了对应分析法。对应分析就克服了上述缺点, 它综合了 R 型和 Q 型因子分析的优点, 并将它们统一起来使得由 R 型的分析结果很容易得到 Q 型的分析结果, 这就克服了 Q 型分析计算量大的困难; 更重要的是可以把变量和样品的载荷反映在相同的公因子轴上, 这样就把变量和样品联系起来便于解释和推断。

对应分析的基本思想是将一个列联表的行和列中各元素的比例结构以点的形式在较低维的空间中表示出来。它的最大特点是能把众多的样品和众多的变量同时做到同一张图解上,将样品的大类及其属性在图上直观而又明了地表示出来,具有直观性。

另外,对应分析还省去了因子选择和因子轴旋转等复杂的数学运算及中间过程,可以从因子载荷图上对样品进行直观的分类,而且能够指示分类的主要参数(主因子),以及分类的依据,是一种直观、简单、方便的多元统计方法。

对应分析法整个处理过程由两部分组成:表格和关联图。对应分析法中的表格是一个二维的表格,由行和列组成。每一行代表事物的一个属性,依次排开。列则代表不同的事物本身,它由样本集合构成,排列顺序并没有特别的要求。在关联图上,各个样本都浓缩为一个点集合,而样本的属性变量在图上同样也是以点集合的形式显示出来的。

设观测值所构成的矩阵为

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

式中, n 为样本观测的次数; p 为变量数,则对应分析方法的基本步骤如下。

由原始观测数据的矩阵 X 出发,计算规格化矩阵 $P = (p_{ij})_{n \times p} = (\frac{x_{ij}}{T})_{n \times p}$ 。

计算过度矩阵 $Z = (z_{ij})_{n \times p}$, 其中, $z_{ij} = \frac{x_{ij} - x_i \cdot x_j / T}{\sqrt{x_i \cdot x_j}}$ 。

进行因子分析,包括 R 型因子分析和 Q 型因子分析。

- R 型因子分析:计算变量协方差矩阵 $A = Z'Z$ 按其累计取前 m 个特征值 $\lambda_1 \quad \lambda_2 \quad \dots \quad \lambda_m$, 并计算相应的单位特征矢量 u_1, u_2, \dots, u_m , 从而得到因子载荷矩阵,即

$$F = \begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & & \vdots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \cdots & u_{pm}\sqrt{\lambda_m} \end{bmatrix}$$

然后在两两因子轴平面上作变量散点图。

- Q 型因子分析:对上述所求的 m 个特征值 $\lambda_1 \quad \lambda_2 \quad \dots \quad \lambda_m$, 计算矩阵 $B = ZZ'D$ 的单位特征矢量 $Zu_1 \approx V_1, Zu_2 \approx V_2, \dots, Zu_m \approx V_m$, 从而得到 Q 型因子载荷矩阵,即

$$G = \begin{bmatrix} V_{11}\sqrt{\lambda_1} & V_{12}\sqrt{\lambda_2} & \cdots & V_{1m}\sqrt{\lambda_m} \\ V_{21}\sqrt{\lambda_1} & V_{22}\sqrt{\lambda_2} & \cdots & V_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & & \vdots \\ V_{n1}\sqrt{\lambda_1} & V_{n2}\sqrt{\lambda_2} & \cdots & V_{nm}\sqrt{\lambda_m} \end{bmatrix}$$

然后在与 R 型相应的因子平面上画样品散点图。

15.2 对应分析

SPSS 中用于分析两个分类变量之间的关系的过程是对应分析 (Correspondence Analysis) 过程。进行多元对应分析的过程是最优标度分析 (Optimal Scaling)。首先介绍对应分析过程。

15.2.1 对应分析过程的参数设置

选择菜单“分析 (Analyze) 降维 (Data Reduction) 对应分析 (Correspondence Analysis)”，则弹出如图 15-1 所示的“对应分析”对话框。



图 15-1 “对应分析 (Correspondence Analysis)”对话框

1. 变量设置

如图 15-1 所示，左边是待分析变量列表。

行 (Row) \ 列 (Column) 栏用于选入对应哪个分析的行变量和列变量。选入变量以后，单击“定义范围 (Define Range)”按钮则弹出如图 15-2 所示对话框。

- 最小值和最大值分别表示相应分类变量的最小值和最大值，此处只可以输入整数，输入后单击“更新 (Update)”按钮则进行确认。
- 类别约束 (Category Constraints) 栏用于设置分类变量取值的约束条件，其下列表框显示的是当前分类变量的取值列表。选中框中的一个值，然后单击右侧的三个单选框设置其约束条件。“无 (None)”表示不作任何的约束；“类别必须相等 (Categories must be Equal)”表示等同约束，表示各类别必须有相同的得分；“类别为补充性 (Categories is Supplemental)”表示增补约束，增补的种类不影响分析过程和种类维数，但会在有效种类的定义空间里被描述。

2. 模型 (Model) 设置

单击图 15-1 中的“模型 (Model)”按钮，则弹出如图 15-3 所示的对话框。

解的维数 (Dimensions in Solution) 栏：指定对应分析的维数。

距离度量 (Distance Measure) 栏：选择行、列分类各自的距离测度。

- 卡方 (Chi-square)： χ^2 距离，系统默认方法。
- 欧氏距离 (Euclidean)。

标准化方法 (Standardization Method) 栏：用于设置标准化的方法，各选项功能如下。

- 除去行列平均值 (Row and column measure removed)：行、列数据都被中心化，当选择 χ^2 距离时，只可以指定该方法。
- 除去行平均值 (Row means are removed)：只有行数据被中心化。
- 除去列平均值 (Column means are removed)：只有列数据被中心化。
- 使行总计相等，并除去平均值 (Row total are equalized and means are removed)：行数据被中心化，且确定中心之前，先令行边际都相等。
- 使列总计相等，并除去平均值 (Column total are equalized and means are removed)：列数据被中心化，且确定中心之前，先令列边际都相等。

正态化方法 (Normalization Method) 栏：此栏用于设置正规化方法，有 5 个选项，各选项功能如下。

- 对称 (Symmetrical)：对称法。
- 行主成分 (Row Principal)：如果要检查行变量内部分类间的距离，选用此方法。
- 主成分 (Principal)：如果要检查行或列变量各自内部分类间的距离，而不是检查行、列间的距离，选用此方法。
- 列主成分 (Column Principal)：如果要检查列变量内部分类间距离，选用此方法。
- 定制 (Custom)：用户自定义，输入一个 -1 ~ 1 之间的数字。-1 相当于 Column Principal 方法，1 相当于 Row Principal 方法，0 相当于 Symmetrical 方法。



图 15-2 “定义范围 (Define Range)”对话框



图 15-3 “模型 (Model) 设置”对话框

3. 统计量 (Statistics) 设置

单击图 15-1 中的“统计量 (Statistics)”按钮，则弹出如图 15-4 所示的对话框，此对话框主要用于设置对应分析输出哪些表格。

- 对应表 (Correspondence Table): 交叉分组列表。
- 行点概览 (Overview of Row Points): 行详细信息表, 包括行变量各分类的得分、质量、惯量、对维度的惯量贡献、维度对分数惯量的贡献。
- 列点概览 (Overview of Column Points): 列详细信息表, 包括列变量各分类的得分、质量、惯量、对维度的惯量贡献、维度对分数惯量的贡献。
- 对应表的排列 (Permutations of the Correspondence Table): 输出按照第一个维度上的得分升序排列的行、列对应表。其下的排列的最大维数 (Maximun dimension for permutations) 可以输入指定表格的最大维数。
- 行概要 (Row Profiles): 表示每个行变量分类对所有列变量分类的分布情况。
- 列概要 (Column Profiles): 表示每个列变量分类对所有行变量分类的分布情况。
- 行点 (Row Point): 表示所有非增补行输出标准差和相关系数。
- 列点 (Column Point): 表示所有非增补列输出标准差和相关系数。

4. 图 (Plots) 设置

单击图 15-1 中的“图 (Plots)”按钮, 则弹出如图 15-5 所示对话框, 此对话框用于设置绘图的选项, 各个选项含义如下。



图 15-4 “统计量 (Statistics) 设置”对话框

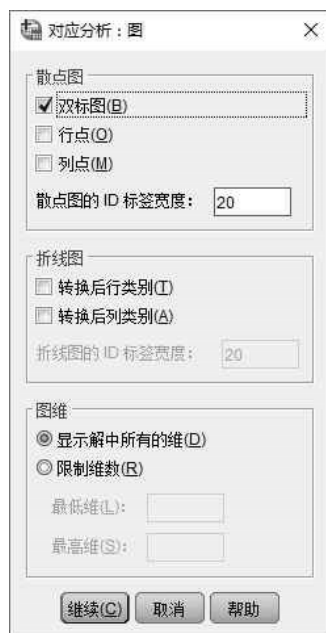


图 15-5 图 (Plots) 对话框

散点图 (Scatterplots): 散点图。

- 双标图 (Biplot): 输出行列的联合分布图。
- 行点 (Row Points): 以矩阵形式输出每个行分类的得分图。
- 列点 (Column Points): 以矩阵形式输出每个列分类的得分图。
- 散点图的 ID 标签宽度 (ID Label Width for Scatterplots): 设置散点图 ID 标签的字符个数, 默认为 20。

折线图 (Line Plots): 线形图。

- 转换后行类别 (Transformed Row Categories): 以行分类的原始取值对行分类的得分作图。
- 转换后列类别 (Transformed Column Categories): 以列分类的原始取值对列分类的得分作图。
- 线图的 ID 标签宽度 (ID Label Width for Line Plots): 设置线形图 ID 标签的字符个数, 默认为 20。

图维数 (Plot Dimensions): 设置输出维度, 对所有输出的多维图形有效。

- 显示解中的所有维数 (Display all dimensions in the solution): 表示分析用到的行列维度都将以交叉矩阵的形式输出。
- 限制维数 (Restrict the number of dimensions): 限制输出指定维度组合的图形, 必须在最低维数 (Lowest) 最高维数 (Highest) 后指定最小、最大维度。

15.2.2 实例分析



结果文件

——附带光盘 “PROGRAM\CH15\实例 15-1 ” 文件夹



动画演示

——附带光盘 “AVI\实例 15-1.avi ” 文件

本实例所用数据为 SPSS 自带的数据集 smoking.sav, 此数据集的格式如图 15-6 所示, 下面就对此数据集进行对应分析。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|-------|----|----|------|-------------|----------------|----|---|----|----|----|
| 1 | staff | 数字 | 4 | 0 | Staff Group | {1, Sr Mana... | 无 | 8 | 右 | 名义 | 输入 |
| 2 | smoke | 数字 | 4 | 0 | Smoking | {1, None}... | 无 | 8 | 右 | 名义 | 输入 |
| 3 | count | 数字 | 4 | 0 | | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 15-6 数据集 smoking.sav 的格式

1. 参数设置

在进行对应分析之前, 需要对数据进行预处理, 选择菜单 “数据 (Data) 个案加权 (Weight Cases)”, 则弹出如图 15-7 所示对话框。选中 “个案加权系数 (Weight cases by)” 选项栏, 则激活其下的 “频率变量 (Frequency Variable)” 选项框, 选中变量 count 到 “频率变量 (Frequency Variable)” 选项框, 然后单击 “确定 (OK)” 按钮返回主界面。

然后进行对应分析的参数设置, 选择菜单 “分析 (Analyze) 降维 (Data Reduction) 对应分析 (Correspondence Analysis)”, 则弹出如图 15-8 所示的 “对应分析” 对话框。选中变量 Staff Group 到 “行 (Row)” 变量框中, 然后单击 “定义范围 (Define Range)” 按钮, 弹出如图 15-9 所示对话框, “最小值 (Minimum Value)” 变量框中填入 1, “最大值 (Maximum

Value)” 变量框中填入 5，并单击“更新 (Update)”按钮确定，最后单击“继续 (Continue)”按钮返回主界面。

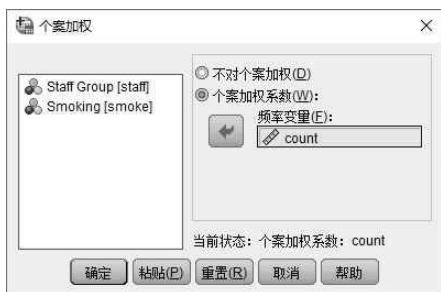


图 15-7 “个案加权 (Weight Cases) 设置”对话框



图 15-8 “对应分析 (Correspondence Analysis)”对话框

然后选择变量 Smoking 到“列 (Column)”变量框中，并单击其下的“定义范围 (Define Range)”按钮，弹出如图 15-10 所示对话框。“最小值 (Minimum Value)”变量框中填入 1，“最大值 (Maximum Value)”变量框中填入 4，并单击“更新 (Update)”按钮确定，最后单击“继续 (Continue)”按钮返回主界面。



图 15-9 “定义范围 (Define Range) 设置”对话框 图 15-10 “定义范围 (Define Range) 设置”对话框

接着单击图 15-8 中的“统计量 (Statistics)”按钮，弹出如图 15-11 所示的对话框，选中“行概要 (Row Profiles)”、“列概要 (Column Profiles)”，以及“对应表的排列 (Permutations of the Correspondence Table)”和“行点 (Row Points)”、“列点 (Column Points)”选项，最后单击“继续 (Continue)”按钮返回主界面。

2. 结果分析

单击主界面“对应分析”对话框 (Correspondence Analysis Dialog Box) 的“确定”按钮，则进行对应分析。首先是图 15-12 的列联表，从图中可以看出不同组别中的人数。



图 15-11 “统计量 (Statistics) 设置”对话框

| 对应表 | | | | | |
|--------------|------|-------|---------|-------|------|
| Staff Group | None | Light | Smoking | | 活动边际 |
| | | | Medium | Heavy | |
| Sr Managers | 4 | 2 | 3 | 2 | 11 |
| Jr Managers | 4 | 3 | 7 | 4 | 18 |
| Sr Employees | 25 | 10 | 12 | 4 | 51 |
| Jr Employees | 18 | 24 | 33 | 13 | 88 |
| Secretaries | 10 | 6 | 7 | 2 | 25 |
| 活动边际 | 61 | 45 | 62 | 25 | 193 |

图 15-12 列联表

如图 15-13 所示的是结果汇总表, 图中给出了 Dimension (维数) Singular Value (奇异值) Intertia (惯量) Chi Square (χ^2 检验), 以及 Sig 数值。

| 摘要 | | | | | | | | |
|----|------|------|--------|-------------------|-------|-------|--------|------|
| 维 | 奇异值 | 惯量 | 卡方 | 显著性 | 惯量比例 | | 置信度奇异值 | |
| | | | | | 占 | 累积 | 标准差 | 相关性 |
| 1 | .273 | .075 | | | .878 | .878 | .070 | .020 |
| 2 | .100 | .010 | | | .118 | .995 | .076 | |
| 3 | .020 | .000 | | | .005 | 1.000 | | |
| 总计 | | .085 | 16.442 | .172 ^a | 1.000 | 1.000 | | |

a. 12 自由度

图 15-13 结果汇总表

如图 15-14 所示为行点概述表, 图中给出行变量的 5 个分组在两个维度中的分值, 其中“质量 (Mass)”表示每组所占的百分比, 后面的 1、2 两项分别为分组在第一维度和第二维度的坐标值, 右侧的“贡献 (Contribution)”项给出了每个分组对各个维度的贡献量, 包括点对维度惯量的贡献和维度对点惯量的贡献。图 15-15 中给出的是列点概述表, 在此不再赘述。

| 行点总览 ^a | | | | | | | | | |
|-------------------|-------|-------|-------|--------|-------|--------|------|------|-------|
| Staff Group | 数量 | 维得分 | | | 惯量 | 贡献 | | | |
| | | 1 | 2 | 点对维的惯量 | | 维对点的惯量 | | 总计 | |
| | | | | 1 | | 2 | 1 | | 2 |
| Sr Managers | .057 | -.126 | .612 | .003 | .003 | .214 | .092 | .800 | .893 |
| Jr Managers | .093 | .495 | .769 | .012 | .084 | .551 | .526 | .465 | .991 |
| Sr Employees | .264 | -.728 | .034 | .038 | .512 | .003 | .999 | .001 | 1.000 |
| Jr Employees | .456 | .446 | -.183 | .026 | .331 | .152 | .942 | .058 | 1.000 |
| Secretaries | .130 | -.385 | -.249 | .006 | .070 | .081 | .865 | .133 | .999 |
| 活动总计 | 1.000 | | | .085 | 1.000 | 1.000 | | | |
| a. 对称正态化 | | | | | | | | | |

图 15-14 行点概述表

| 列点总览 ^a | | | | | | | | | |
|-------------------|-------|-------|-------|------|--------|-------|------|------|-------|
| Smoking | 数量 | 维得分 | | 惯量 | 点对维的惯量 | | 贡献 | | 总计 |
| | | 1 | 2 | | 1 | 2 | 1 | 2 | |
| None | .316 | -.752 | .096 | .049 | .654 | .029 | .994 | .006 | 1.000 |
| Light | .233 | .190 | -.446 | .007 | .031 | .463 | .327 | .657 | .984 |
| Medium | .321 | .375 | -.023 | .013 | .166 | .002 | .982 | .001 | .983 |
| Heavy | .130 | .562 | .625 | .016 | .150 | .506 | .684 | .310 | .995 |
| 活动总计 | 1.000 | | | .085 | 1.000 | 1.000 | | | |

a. 对称正态化

图 15-15 列点概述值

图 15-16 是对应分析图，从这个图中可以看出所需要的分析结果。



图 15-16 对应分析图

15.3 最优标度过程

15.3.1 最优标度过程的参数设置

选择菜单“分析 (Analyze) 降维 (Data Reduction) 最优标度 (Optimal Scaling)”，则系统弹出如图 15-17 所示对话框，此对话框用于选择采用何种最优标度分析方法。

最佳度量级别 (Optimal Scaling Level) 栏，设置变量的度量类型。

- 所有变量均为多重名义 (All Variables are Multiple Nominal): 适合于所有变量均为无序多分类。
- 某些变量并非多重名义 (Some Variable (s) are not Multiple Nominal): 适合于有的变量是单分类的名义变量、有序分类变量或者离散的数值型变量。

变量集数目 (Number of Sets of Variables) 栏, 用于设置变量集的个数。

- 一个集合 (One Set): 表示只分析一组变量间的关系。
- 多个集合 (Multiple Sets): 适用于数据集中存在多选题变量集, 即有多个变量是同属于一道多选题的不同答案。

选定的分析 (Selected Analysis) 栏, 用于显示当前选项所使用的分析方法。

- 多重对应分析 (Multiple Correspondence Analysis): 多元对应分析。
- 分类主要成分 (Categorical Principal Components): 分类变量的主成分分析法。
- 非线性典型相关性 (Nonlinear Canonical Correlation): 非线性典型相关方法。



图 15-17 最优尺度分析方法选择

单击图 15-17 中的“定义 (Define)”按钮, 则弹出如图 15-18 所示的“多重对应分析设置”对话框。此界面可以设置有关多元对应分析的各个参数属性, 具体设置如下所述。

1. 变量选择设置

图 15-18 的左边是待分析的变量列表, 各个列表框功能如下所述。

- 分析变量 (Analysis Variables): 选入分析变量, 选入后则会激活“定义变量权重 (Define Variable Weight)”按钮, 单击此按钮, 弹出如图所示 15-19 所示的对话框。可以设置变量权重, 默认为 1。
- 补充变量 (Supplementary Variable): 增补变量列表。
- 标记变量 (Labeling Variables)
- 解的维数 (Dimension in Solution): 设置描述分析结果的低维空间维数, 默认为 2。

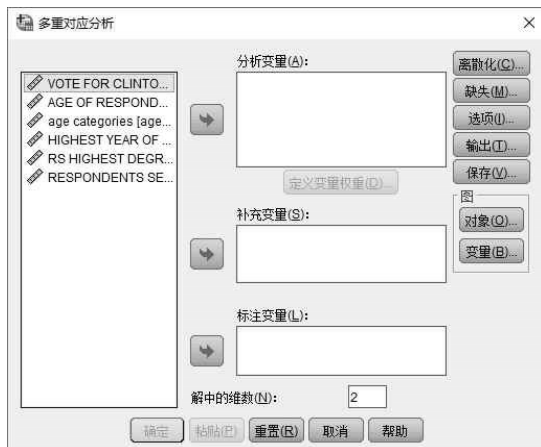


图 15-18 “多重对应分析 (Multiple Correspondence Analysis) 设置”对话框



图 15-19 “定义权重设置 (Define Variable Weight)”对话框

2. 离散化 (Discretize) 设置

单击图 15-18 中的“离散化 (Discretize)”按钮，则弹出如图 15-20 所示的对话框，即“离散化设置”对话框，各选项框含义如下所述。

变量 (Variables) 栏：变量名列表，括号中是对应的离散化方法。

方法 (Method) 下拉菜单：设置离散化方法，选择后单击“更改 (Change)”按钮确定改变，如图 15-21 所示。

- 未指定 (Unspecified)：无离散化操作。
- 分组 (Grouping)：将取值重新编码为固定个数或者固定间隔的类别。
- 排秩 (Ranking)：将变量取值后排序，取其秩进行分类。
- 乘 (Multiplying)：将当前值标准化后乘以 10，再取整，最后加上一个常数，使得离散化后的最小值为 1。

分组 (Grouping) 栏：此栏用于设置选定分组 (Grouping) 方法后的一些参数。



图 15-20 “离散化 (Discretize) 设置”对话框

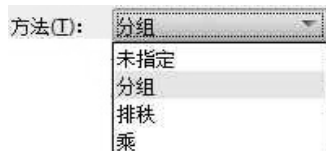


图 15-21 离散化方法

3. 缺失 (Missing) 设置

缺失 (Missing) 设置即为缺失值设置，单击图 15-18 中的“缺失 (Missing)”按钮，则弹出如图 15-22 所示对话框。

(1) 缺失值策略 (Missing Value Strategy) 栏

此栏用于设置变量，其中“分析变量 (Analysis Variables)”栏用于显示当前选入的分析变量；“补充变量 (Supplementary Variable)”栏用于显示当前选入的增补变量列表，变量名后括号显示的是对应的缺失值处理方法。

(2) 策略 (Strategy) 栏

此栏设置缺失值处理方法。

- 排除缺失值 (Exclude Missing Values)：排除含有缺失值的变量，“众数 (Mode)”表示用类别取值的众数来取代缺失值，如存在多个众数，则取指标最小的；“附加类别 (Extra category)”表示用一个额外的分类值取代所有缺失值。

- 插补缺失值 (Impute Missing Values): 缺失值的替代。
- 排除对于此变量具有缺失值的对象 (Exclude objects with missing values on this variables): 排除分析变量含缺失值的观测, 此方法对增补变量无效。

4. 选项 (Options) 设置

单击图 15-18 中的“选项 (Options)”按钮, 则弹出如图 15-23 所示对话框, 单击“继续 (Continue)”按钮则返回主界面。

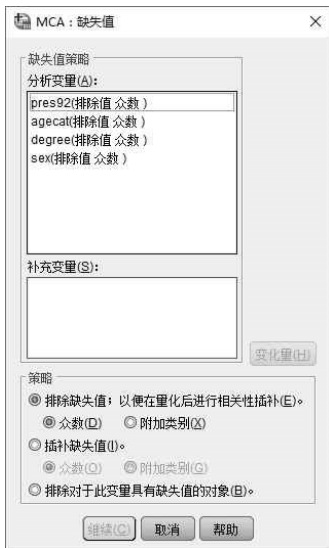


图 15-22 “缺失值 (Missing) 设置”对话框



图 15-23 “选项 (Options) 设置”对话框

补充对象 (Supplementary Objects) 栏, 设置增补观测在数据集里的记录号, 可以通过“更改 (Change)”、“除去 (Remove)”按钮进行更改或删除。

- 个案范围 (Range of Cases): 指定第一个 (First) 和最后一个 (Last), 然后单击“添加 (Add)”按钮加入到下面的增补列表。
- 单个个案 (Single Case): 设置特定的行号, 单击“添加 (Add)”按钮加入增补列表中。

正态化方法 (Normalization Method) 栏, 设置变量或者观测得分的正态化方法, 如图 15-24 所示下拉菜单。

- 变量主成分 (Variable Principal): 相当于简单对应分析 Column Principal 方法。
- 对象主成分 (Object Principal): 相当于简单对应分析的 Row Principal 方法。
- 对称 (Symmetrical): 相当于简单对应分析的 Symmetrical 方法。
- 独立 (Independent): 相当于简单对应分析的 Principal 方法。
- 定制 (Custom): 相当于简单对应分析的 Custom 方法。

条件 (Criteria) 栏, 用来设置模型的拟合标准。“收敛 (Convergence)”用来指定收敛的临界值; “最大迭代次数 (Maximum Iterations)”用来指定的循环次数。

图的标注依据 (Label Plots By) 栏, 用来设置输出图形的显示方式。

- 变量标签或值标签 (Variable labels or value labels): 显示变量标签或变量值, 其下的标签长度的限制 (Limit for label length) 用来指定变量标签的最大长度。

- 变量名称或值 (Variable name or values) : 显示变量名或变量值。
 图维数 (Plot Dimensions) 栏, 设置输出图形的维数, 与图 15-5 的设置方法一样。
 配置 (Configuration) 栏, 用来设置从一个文件读入坐标的结构信息, 单击“文件 (File)”按钮则选择文件。

5. 输出 (Output) 设置

单击图 15-18 中的“输出 (Output)”按钮, 则弹出如图 15-25 所示的对话框, 各个选项栏功能如下所述。



图 15-24 “正态化方法 (Normalization Method)
选择”对话框

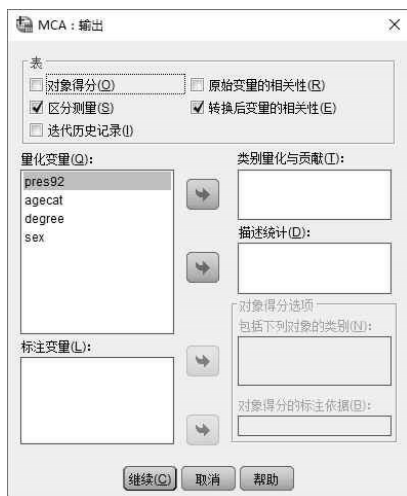


图 15-25 “输出 (Output) 设置”
对话框

表 (Tables) 栏 : 此栏用于选择输出的表格。

- 对象得分 (Object Scores) : 观测得分表。
- 原始变量的相关性 (Correlations of Original Variables) : 初始变量取值的相关系数矩阵及其特征值。
- 区分测量 (Discrimination Measures) : 输出每个变量、维度的判别度量方式。
- 替换后变量的相关性 (Correlations of Transformed Variables) : 输出变换后变量的相关系数矩阵及其特征值。
- 迭代历史记录 (Iteration History) : 输出迭代过程中方差的变化过程。

类别量化与贡献 (Category Quantifications and Contributions) : 量化变量列表, 对其每一维度输出类别量化的信息。

描述统计量 (Descriptive Statistics) : 描述变量列表, 输出其频数, 缺失值个数, 众数等基本统计信息。

6. 保存 (Save) 设置

单击图 15-18 中的“保存 (Save)”按钮, 则弹出如图 15-26 所示的对话框, 各保存选项的功能如下所述。

离散化数据 (Discretized Data) 栏: 此栏用于设置离散化数据的选项。

- 创建离散化数据 (Create Discretized Data): 创建一个保存离散数据的数据集。
- 创建新数据集 (Create a New Dataset): 表示建立一个新的数据集来保存指定数据。

对象得分 (Object Scores) 栏: 此栏用于设置保存观测得分数据选项。

- 将对象得分保存到活动数据集 (Save object scores to the active dataset): 把指定数据存入到当前数据集中。
- 创建新数据集 (Create a New Dataset): 表示建立一个新的数据集来保存指定数据。
- 写入新数据文件 (Write a New Data File): 表示建立一个新的文件类保存指定数据, 单击“文件 (File)”按钮指定文件。

转换后变量 (Transformed Variables) 栏: 设置保存变量变换数据的选项。

- 将转换后变量保存到活动数据集 (Save transformed variables the active dataset): 把指定数据存入到当前数据集中。
- 创建新数据集 (Create a New Dataset): 表示建立一个新的数据集来保存指定数据。
- 写入新数据文件 (Write a New Data File): 表示建立一个新的文件类保存指定数据, 单击“文件 (File)”按钮指定文件。



图 15-26 “保存 (Save) 设置”对话框

7. 对象 (Object) 设置

单击图 15-18 中的“对象 (Object)”按钮, 则会弹出如图 15-27 所示的对话框。

图 (Plots) 栏, 用来设置作图类别。

- 对象点 (Object Points): 只对对象点作图。
- 对象和质心 (双标图) (Objects and Centroids (biplot)): 对对象点及其中心点作图。

双标图变量 (Biplot Variables) 栏, 此栏用于设置行、列联合分数图的变量。

- 所有变量 (All Variables): 使用全部变量。
- 选定变量 (Selected Variables): 把需要的变量从“可用 (Avaliable)”列表中选入

“选定 (Selected)” 列表中。

标注对象 (Label Objects) 栏，此栏用于设置标识对象的标签变量。

- 个案号 (Case Number): 以行号作为标签。
- 变量 (Variable): 把标签变量从“可用 (Avaliable)”列表中选入“选定 (Selected)”列表”中。

8. 变量 (Variable) 设置

单击图 15-18 中的“变量 (Variable)”按钮，则弹出如图 15-28 所示对话框，各个选项功能如下所述。

类别图 (Category Plots) 栏：此栏对选入的变量绘制图形。

联合类别图 (Joint Category Plots) 栏：在每一个图形之中显示所有选入变量各类别的中心值。

转换图 (Transformation Plots) 栏：对选入的变量绘制“最优量化值”和“类别指示变量”的相关图形。“维数 (Dimensions)”中输入维数，每个维数输出一个图形；“包含残差图 (Include Residual Plots)”表示绘制变量的输出残差图。

区分测量 (Discrimination Measures) 栏。

- 显示图 (Display Plot): 为指定变量输出区分度量的图形。
- 使用所有变量 (Use all Variables): 表示使用全部变量。
- 使用选定变量 (Use Selected Variables): 把需要使用的变量从坐标的变量列表框中选入到其下的列表框中。



图 15-27 “对象 (Object) 作图设置”对话框

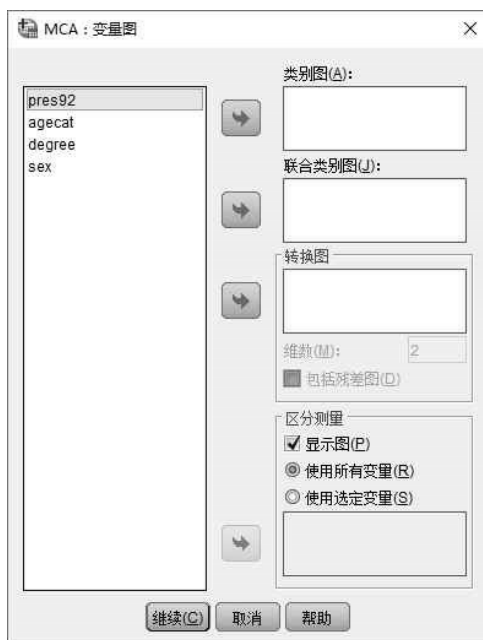


图 15-28 “变量 (Variable) 设置”对话框

15.3.2 实例分析



结果文件——附带光盘“PROGRAM\CH15\实例 15-2”文件夹



动画演示——附带光盘“AVI\实例 15-2.avi”文件

本实例选择 SPSS 自带的文件 voter.sav，此数据集是关于美国总统选举的调查数据集，包括 6 个变量，即 pres92（候选人）、age（年龄）、agecat（年龄分类组别）、educ（教育）、degree（学位）、sex（性别）。数据集的格式如图 15-29 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|--------|----|----|------|-----------------|------------------|------------|---|----|----|----|
| 1 | pres92 | 数字 | 1 | 0 | VOTE FOR CLI... | {1, Bush}... | 4 - 9, 0 | 8 | 右 | 标度 | 输入 |
| 2 | age | 数字 | 2 | 0 | AGE OF RESP... | 无 | 0, 98, 99 | 8 | 右 | 标度 | 输入 |
| 3 | agecat | 数字 | 8 | 2 | age categories | {1.00, lt 35}... | 无 | 8 | 右 | 标度 | 输入 |
| 4 | educ | 数字 | 2 | 0 | HIGHEST YEA... | 无 | 97, 98, 99 | 8 | 右 | 标度 | 输入 |
| 5 | degree | 数字 | 1 | 0 | RS HIGHEST D... | {0, lt high s... | 7, 8, 9 | 8 | 右 | 标度 | 输入 |
| 6 | sex | 数字 | 1 | 0 | RESPONDENT... | {1, male}... | 无 | 8 | 右 | 标度 | 输入 |

图 15-29 数据文件 voter.sav 的格式

1. 参数设置

依次选择菜单“分析（Analyze）”→“降维（Data Reduction）”→“最优标度（Optimal Scaling）”，则系统弹出如图 15-30 所示对话框，此对话框用于选择采用何种最优标度分析方法。选择“所有变量均为多重标称（All Variables Multiple Nominal）”选项栏和“一个集合（One Set）”变量栏。

然后单击“定义（Define）”按钮，弹出如图 15-31 所示对话框，选择变量 pres92（候选人）、agecat（年龄分类组别）、degree（学位）、sex（性别）到“分析变量（Analysis Variables）”选项栏中。然后单击“定义变量权重（Define Variable Weight）”按钮，弹出如图 15-32 所示对话框，在“变量权重（Variable Weight）”变量框中填入 1，然后单击“继续（Continue）”按钮返回主界面。

单击图 15-31 中的“离散化（Discretize）”按钮，则弹出如图 15-33 所示对话框。

单击图 15-31 中的“输出（Output）”按钮，则弹出如图 15-34 所示对话框，选中“原始变量的相关性（Correlation of original variables）”选项栏，选择变量 pres92 到“类别量化与贡献（Category Quantifications and Contributions）”选项栏中，选择变量 sex 到“描述统计量



图 15-30 “最优尺度（Optimal Scaling）设置”对话框

(Descriptive Statistics)”变量框中。然后单击“继续 (Continue)”按钮返回主界面。

单击图 15-31 中的“对象 (Object)”按钮，则弹出如图 15-35 所示对话框，选中“对象点 (Object Points)”选项栏，然后单击“继续 (Continue)”按钮返回主界面。

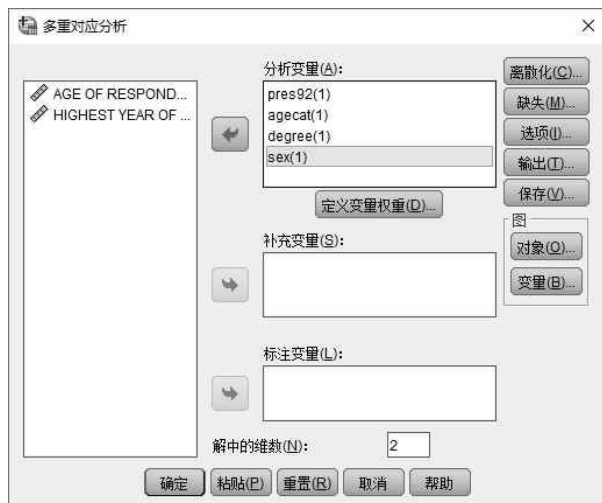


图 15-31 “最优标度 (Optimal Scaling) 设置”对话框



图 15-32 定义变量权重 (Define Variable Weight) 设置对话框



图 15-33 “离散化 (Discretize)”对话框



图 15-34 “输出 (Output) 设置”对话框

单击图 15-31 中的“变量 (Variable)”按钮，则弹出如图 15-36 所示对话框，选中变量 degree 到“类别图 (Category Plots)”变量框中，选择变量 pres92、agecat、degree、sex 到“联合类别图 (Joint Category Plots)”变量框中，然后单击“继续 (Continue)”按钮返回主界面。



图 15-35 “对象 (Object) 设置”对话框



图 15-36 “变量 (Variable) 设置”对话框

2. 结果分析

设置好各种参数以后，单击主界面的“确定”按钮进行对应分析，结果如下。首先是如图 15-37 所示的描述性统计量。包括本案例的一些统计信息，以及被调查者的性别统计信息。

| 个案处理摘要 | |
|-------------------------|------|
| 有效活动个案 | 1658 |
| 具有缺失值的活动个案 ^a | 189 |
| 补充个案 | 0 |
| 总计 | 1847 |
| 在分析中使用的个案 | 1847 |

a. 出现小于或等于零的值 (请参阅“警告”表)。

| RESPONDENTS SEX | | |
|-----------------|---------------------|------|
| 频率 | | |
| 有效 | male | 804 |
| | female ^a | 1043 |
| | 总计 | 1847 |

a. 众数。

图 15-37 描述性统计量

然后是模型信息，如图 15-38 所示，表中给出了两个维度的方差总计及其惯量信息。

| 模型摘要 | | | |
|------|-------------------|----------|------|
| 维 | 克隆巴赫 Alpha | 方差所占百分比 | |
| | | 总计 (特征值) | 惯量 |
| 1 | .329 | 1.327 | .332 |
| 2 | .215 | 1.192 | .298 |
| 总计 | | 2.519 | .630 |
| 平均值 | .275 ^a | 1.260 | .315 |

a. 克隆巴赫 Alpha 平均值基于平均特征值。

图 15-38 模型信息

接着输出的是类别点联合图形，如图 15-39 所示，此图是把四个分析变量的类别点中心坐标在一个图形中加以显示的效果。

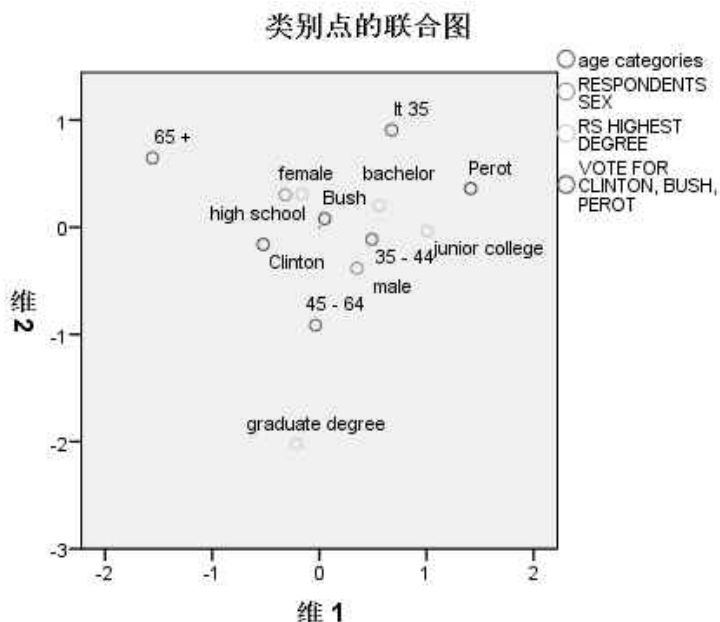


图 15-39 类别点联合图形

如图 15-40 所示为区分度量图形，区分度量反映了维度得分和量化后变量值的相关性大小，由此图可以判断重点变量在其相关性较大的维度上的特征，在这个维度上的类别点一般会分得更开。由图中可以看到学历高低在维度 2 上值得受较大关注；年龄段在两个维度上均需要关注；性别变量的区分度量在两个维度都很小。

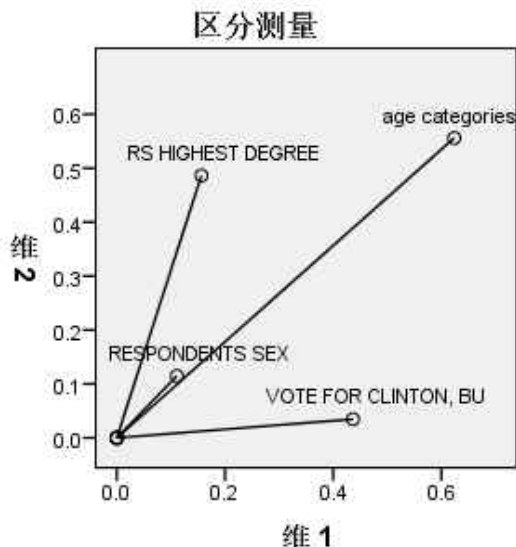
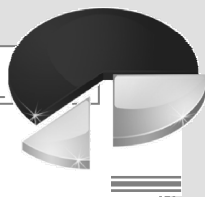


图 15-40 区分度量图形



第 16 章 可靠性和多维标度分析

可靠性是指一个衡量的正确性或精确性,可靠性包括稳定性及一致性;学者 Kerlinger 认为可靠性可以衡量出工具(问卷)的可靠度、一致性与稳定性。

一般可靠性的测量容易产生误差的原因,来自研究者的因素包括测量内容(遣词造句、问题形式等)不当、情境(时间长短、气氛、前言说明等),以及研究者本身的疏忽(听错、记错等);而来自受访者的因素则可能是由于其个性、年龄、教育程度、社会阶层及其他心理因素等,而影响其答题的正确性。

问卷内容的同构型及受访时间间隔的影响是影响可靠性的两个主要因素。

研究者透过可靠性与效度的检验,可以了解测量工具问卷本身是否优良适当,以作为改善修正的根据,并可避免做出错误的判断。

本章将叙述可靠性分析在 SPSS 中的实现,以及实例分析。



本讲内容

- 可靠性分析
- 多维标度分析

16.1 可靠性分析

16.1.1 可靠性分析的基本原理

可靠性 (Reliability), 是指采用同样的方法对同一对象重复测量时所得结果的一致性程度。可靠性指标多以相关系数表示,大致可分为三类:稳定系数(跨时间的一致性),等值系数(跨形式的一致性)和内在一致性系数(跨项目的一致性)。

作好问卷调查后,接下来为了进一步考验问卷的可靠性与有效性,即要做可靠性分析 (Reliability Analysis),可靠性本身与测量所得结果正确与否无关,它的功用在于检验测量本身是否稳定。

测验可靠性越高,表示测验结果越可信,但也无法期望两次测验结果完全一致,可靠性除受测验质量影响外,也受很多其他受测者因素的影响,故没有一份测验是完全可靠

的。可靠性只是一种程度上大小的差别。一致性高的问卷便是指同一群人接受性质相同、题型相同、目的相同的各种问卷测量后，在各衡量结果间显示出强烈的正相关。稳定性高的测量工具则是指一群人在不同时空下接受同样的衡量工具时，结果的差异很小。

可靠性分析的方法主要有以下四种。

1. 重测可靠性法

这一方法是用同样的问卷对同一组被调查者间隔一定时间重复施测，计算两次施测结果的相关系数。显然，重测可靠性属于稳定系数。重测可靠性法特别适用于事实式问卷，如性别、出生年月等在两次施测中不应有任何差异，大多数被调查者的兴趣、爱好、习惯等在短时间内也不会有十分明显的变化。如果没有突发事件导致被调查者的态度、意见突变，这种方法也适用于态度、意见式问卷。由于重测可靠性法需要对同一样本试测两次，被调查者容易受到各种事件、活动和他人的影响，而且间隔时间长短也有一定限制，因此在实施中有一定困难。

2. 复本可靠性法

复本可靠性法是让同一组被调查者一次填答两份问卷复本，计算两个复本的相关系数。复本可靠性属于等值系数。复本可靠性法要求两个复本除表述方式不同外，在内容、格式、难度和对应题项的提问方向等方面要完全一致，而在实际调查中，很难使调查问卷达到这种要求，因此采用这种方法者较少。

3. 折半可靠性法

折半可靠性法是将调查项目分为两半，计算两半得分的相关系数，进而估计整个量表的可靠性。折半可靠性属于内在一致性系数，测量的是两半题项得分间的一致性。这种方法一般不适用于事实式问卷（如年龄与性别无法相比），常用于态度、意见式问卷的可靠性分析。在问卷调查中，态度测量最常见的形式是 5 级李克特（Likert）量表。进行折半可靠性分析时，如果量表中含有反意题项，应先将反意题项的得分作逆向处理，以保证各题项得分方向的一致性，然后将全部题项按奇偶或前后分为尽可能相等的两半，计算二者的相关系数（ r_{hh} ，即半个量表的可靠性系数），最后用斯皮尔曼-布朗（Spearman-Brown）公式求出整个量表的可靠性系数（ r_u ）。

4. α 可靠性系数法

Cronbach α 可靠性系数是目前最常用的可靠性系数，其公式为

$$\alpha = \frac{K}{K-1} \left[1 - \frac{\sum S_i^2}{S_X^2} \right]$$

式中， K 是量表中题项的总数， S_i 是第 i 题得分的题内方差， S_X 是全部题项总得分的方差。从公式中可以看出， α 系数评价的是量表中各题项得分间的一致性，属于内在一致性系数。这种方法适用于态度、意见式问卷（量表）的可靠性分析。

可靠性高低与 Cronbach α 可靠性系数相互对照参见表 16-1。

表 16-1 可靠性高低对照表


| 可 靠 性 | Cronbach α 系数 |
|----------|-------------------------|
| 不可信 | $\alpha < 0.3$ |
| 勉强可信 | $0.3 \leq \alpha < 0.4$ |
| 可信 | $0.4 \leq \alpha < 0.5$ |
| 很可信(最常见) | $0.5 \leq \alpha < 0.7$ |
| 很可信(次常见) | $0.7 \leq \alpha < 0.9$ |
| 十分可信 | $0.9 \leq \alpha$ |

测量一组同义或平行测验总和的可靠性,如果尺度中的所有项目都在反映相同的特质,则各项目之间应具有真实的相关存在。若某一项目和尺度中其他项目之间并无相关存在,就表示该项目不属于该尺度,而应将之剔除。

16.1.2 可靠性分析的参数设置

选择菜单“分析(Analyze) 标度(Scale) 可靠性分析(Reliability Analysis)”,即进入“可靠性分析”对话框,如图 16-1 所示。各个设置选项功能如下所述。

1. 选择分析变量

如图 16-1 所示,左边的框中所示的是待分析的对话框,右边是需要分析的变量框,选中左边框中的变量,单击按钮即可被选入分析的变量列表之中。

项目(Items)框:用于选入分析的变量。

模型(Model)选项:此选项用于指定要使用的可靠性系数,如图 16-2 所示。



图 16-1 “可靠性分析(Reliability Analysis)设置”对话框



图 16-2 Model 选项

- Alpha: 表示 Cronbach α 系数,系统默认选择项。
 - 折半(Split-half): 表示分半可靠性。
 - 格特曼: 表示 Guttman 系数,输出的 Lambda3 实际就是 Cronbach α 系数。
 - 平行(Parallel): 表示平行测验的可靠性估计。
 - 严格平行(Strict Parallel): 在平行测验的基础上,要求各变量的均值相等。
- 刻度标签(Scale Label): 指定刻度标签框。

2. 统计量设置

单击图 16-1 中的“统计量 (Statistics)”按钮, 则弹出如图 16-3 所示的“统计量设置”对话框。各个选项栏功能如下。

描述性 (Descriptives for) 栏: 此栏主要是选择要输出的统计量。

- 项 (Item): 输出各变量的均值、标准差等统计信息。
- 标度 (Scale): 输出各变量之和的均值、方差、标准差等信息。
- 删除项后的标度 (Scale if item deleted): 输出问卷中删除指定变量后, 相应统计量的改变值。

摘要 (Summaries) 栏: 此栏设置关于各项目的描述统计量的输出。

- 平均值 (Means): 输出项目均数的最小、最大、平均值, 项目均数的极差、方差, 最大项目均数与最小项目均数之比。
- 方差 (Variances): 输出项目方差的最小、最大、平均值, 项目方差的极差和方差, 最大项目方差与最小项目方差之比。
- 协方差 (Covariances): 输出项目协方差的最小、最大、平均值, 项目协方差的极差和方差, 项目协方差的最大项和最小项之比。
- 相关性 (Correlations): 输出项目相关系数的最小、最大、平均值, 项目相关系数的极差和方差, 项目相关系数的最大项和最小项之比。

项之间 (Inter-Item) 栏: 此栏设置输出变量间的相关信息。

- 相关性 (Correlations)
- 协方差 (Covariances)

ANOVA 表 (ANOVA Table) 栏: 此栏主要设置方差分析选项。

- 无 (None): 不进行分析。
- F 检验 (F_test): 相当于重复测量的方差分析, 该方法适用于数据呈正态分布的情况。
- 傅莱德曼卡方 (Friedman chi-square): 输出 Friedman χ^2 统计量和 Kendall 调谐系数, 该方法适用于取秩格式的数据, 可以取代方差分析中的 F 检验。
- 柯克兰卡方 (Cochran chi-square): 对各变量进行 Cochran χ^2 检验。

霍特林 T 平方: 进行多元检验。

图基可加性检验: 检验各变量之间是否具有显著的交互作用。

同类相关系数 (Intraclass correlation coefficient) 栏: 此栏设置关于组内相关系数的选项。

- 模型 (Model): 此下拉菜单用于指定计算组内相关系数的模型, 如图 16-4 所示。其中“双向混合 (Two-Way Mixed)”为两方向固定模型; “双向随机 (Two-Way Random)”为两方向随机模型; “单向随机 (One-Way Random)”为单方向随机模型。
- 类型 (Type): 此下拉菜单用于指定指标的类型, 可选择“一致性 (Consistency)”和“绝对一致性 (Absolute Agreement)”。
- 置信区间 (Confidence Interval): 用于指定置信区间。
- 检验值 (Test Value): 用于指定一个用于和观测相关系数进行比较的待检验数值, 输入数值要求为 0~1, 系统默认为 0。



图 16-3 “统计量 (Statistics) 设置”对话框

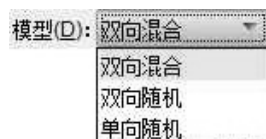




图 16-4 “模型 (Model) 选项”菜单

16.1.3 实例分析

-  **结果文件** —— 附带光盘 “PROGRAM\CH16\实例 16-1” 文件夹
-  **动画演示** —— 附带光盘 “AVI\实例 16-1.avi” 文件

本实例所用数据为 SPSS 中自带的数据集 tv-survey.sav，此数据集是某个电视台对它们电视节目收视用户的调查，下面对此调查数据进行可靠性分析，数据集 tv-survey.sav 的数据格式如图 16-5 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|----------|----|----|------|------------------------|------------|----|---|----|----|----|
| 1 | any | 数字 | 1 | 0 | Any reason | {0, NO}... | 无 | 8 | 右 | 有序 | 输入 |
| 2 | bored | 数字 | 1 | 0 | No other popula... | {0, NO}... | 无 | 8 | 右 | 标度 | 输入 |
| 3 | critics | 数字 | 1 | 0 | Critics still give ... | {0, NO}... | 无 | 8 | 右 | 标度 | 输入 |
| 4 | peers | 数字 | 1 | 0 | Other people st... | {0, NO}... | 无 | 8 | 右 | 标度 | 输入 |
| 5 | writers | 数字 | 1 | 0 | The original scr... | {0, NO}... | 无 | 8 | 右 | 标度 | 输入 |
| 6 | director | 数字 | 1 | 0 | The original dire... | {0, NO}... | 无 | 8 | 右 | 标度 | 输入 |
| 7 | cast | 数字 | 1 | 0 | The original cas... | {0, NO}... | 无 | 8 | 右 | 标度 | 输入 |

图 16-5 数据集 tv-survey.sav 的数据格式

1. 参数设置

选择菜单“分析 (Analyze) 标度 (Scale) 可靠性分析 (Reliability Analysis)”，即进

入“可靠性分析”对话框,如图 16-6 所示。各个设置选项功能如下所述。把所有变量均选入“项目 (Items)”变量框中。

然后单击图 16-6 中的“统计量 (Statistics)”按钮,弹出如图 16-7 所示对话框,选择“项 (Item)”选项栏,选择“相关性 (Correlations)”选项栏,然后单击“继续 (Continue)”按钮返回主界面。



图 16-6 “可靠性分析设置”对话框

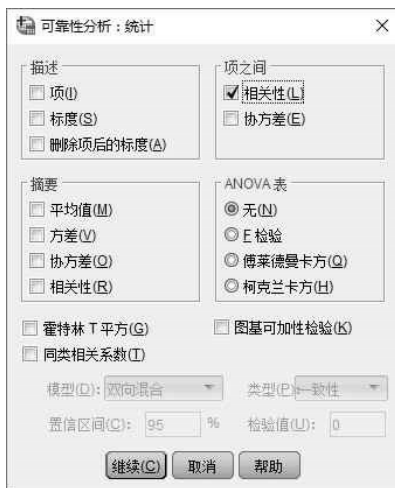


图 16-7 “统计量设置”对话框

2. 结果分析

设置好上述参数以后,则单击可靠性分析对话框 (Reliability Analysis dialog box) 主界面的“确定”按钮进行可靠性分析。结果如下。首先是基本的个案处理结果,如图 16-8 所示。给出了基本的统计信息,有效观测数量为 906,没有缺失值。

| 个案处理摘要 | | | |
|--------|-----------------|-----|-------|
| | | 个案数 | % |
| 个案 | 有效 | 906 | 100.0 |
| | 排除 ^a | 0 | .0 |
| | 总计 | 906 | 100.0 |

a. 基于过程中所有变量的成列删除。

图 16-8 个案处理结果

然后输出的是可靠性统计量,如图 16-9 所示。此表给出了 Cronbach α 可靠性系数计算结果。Cronbach α 可靠性系数为 0.898,可见此问卷的可靠性比较理想。

| 可靠性统计 | | |
|------------|-------------------|----|
| 克隆巴赫 Alpha | 基于标准化项的克隆巴赫 Alpha | 项数 |
| .898 | .894 | 7 |

图 16-9 可靠性统计量

如图 16-10 所示的是各变量统计信息,包括均值、标准差等信息。

最后输出的是相关矩阵,如图 16-11 所示,前四项之间的相关性较高,说明大部分用户是基于前四种原因观看节目的。

| 项统计 | | | |
|--|-----|------|-----|
| | 平均值 | 标准差 | 个案数 |
| Any reason | .49 | .500 | 906 |
| No other popular shows on at that time | .50 | .500 | 906 |
| Critics still give the show good reviews | .50 | .500 | 906 |
| Other people still watch the show | .53 | .499 | 906 |
| The original screenwriters stay | .81 | .389 | 906 |
| The original directors stay | .83 | .378 | 906 |
| The original cast stays | .89 | .315 | 906 |

图 16-10 各变量统计信息

| 项间相关性矩阵 | | | | | | | |
|--|------------|--|--|-----------------------------------|---------------------------------|-----------------------------|-------------------------|
| | Any reason | No other popular shows on at that time | Critics still give the show good reviews | Other people still watch the show | The original screenwriters stay | The original directors stay | The original cast stays |
| Any reason | 1.000 | .815 | .813 | .782 | .408 | .421 | .303 |
| No other popular shows on at that time | .815 | 1.000 | .826 | .807 | .422 | .423 | .307 |
| Critics still give the show good reviews | .813 | .826 | 1.000 | .804 | .458 | .453 | .336 |
| Other people still watch the show | .782 | .807 | .804 | 1.000 | .443 | .460 | .340 |
| The original screenwriters stay | .408 | .422 | .458 | .443 | 1.000 | .632 | .625 |
| The original directors stay | .421 | .423 | .453 | .460 | .632 | 1.000 | .600 |
| The original cast stays | .303 | .307 | .336 | .340 | .625 | .600 | 1.000 |

图 16-11 相关矩阵

16.2 多维标度分析

16.2.1 多维标度分析简介

多维标度分析是市场调查、分析数据的统计方法之一。通过多维标度分析,可以将消费者对商品相似性的判断产生一张能够看出这些商品间相关性的图形。

例如,有十个百货商场,让消费者排列出对这些百货商场两两间相似的感知程度,根据这些数据,用多维标度分析,可以判断消费者认为哪些商场是相似的,从而可以判断竞争对手。

用于反映多个研究事物间相似(不相似)程度,通过适当的降维方法,将这种相似(不相似)程度在低维度空间中用点与点之间的距离表示出来,并有可能帮助识别那些影响

事物间相似性的潜在因素。这种方法在市场研究中应用得非常广泛。

它使用的数据是消费者对一些商品相似程度（或差异程度）的评分，通过分析产生一张能够看出这些商品间相关性的图形（感知图）。

由于多维标度分析法通常是基于研究对象之间的相似性（距离）的，只要获得了两个研究对象之间的距离矩阵，就可以通过相应统计软件做出它们的相似性知觉图。


在实际应用中，距离矩阵的获得主要有两种方法：一种是采用直接的相似性评价，先将所有评价对象进行两两组合，然后要求被访者对所有的这些组合间进行直接相似性评价，这种方法称为直接评价法；另一种为间接评价法，由研究人员根据事先经验，找出影响人们评价研究对象相似性的主要属性，然后对每个研究对象，让被访者对这些属性进行逐一评价，最后将所有属性作为多维空间的坐标，通过距离变换计算对象之间的距离。

多维标度分析的主要思路是利用对被访者对研究对象的分组，来反映被访者对研究对象相似性的感知，这种方法具有一定直观合理性。同时该方法实施方便，调查中被访者负担较小，很容易得到理解并且接受。当然，该方法的不足之处是牺牲了个体距离矩阵，由于每个被访者个体的距离矩阵只包含 1 与 0 两种取值，相对较为粗糙，个体距离矩阵的分析显得比较勉强。但这一点是完全可以接受的，因为对大多数研究而言，并不需要知道每一个体的空间知觉图。

16.2.2 多维标度过程的参数设置

SPSS 中用于多维标度分析的是 ALSCAL 过程，选择菜单“分析（Analyze）标度（Scale）多维标度（Multidimensional Scaling）”，则弹出“进行多维标度分析”对话框，如图 16-12 所示，各个选择项的具体功能如下所述。

1. 变量设置

如图 16-12 所示的左边变量框是待分析的变量框，选中变量后，单击按钮  即可选入变量（Variables）列表框。

变量（Variables）列表框，用于选入表示距离的分析变量。

个别矩阵（Individual Matrices for）框，用于选入分组变量，分析时将会为每一组变量分别计算距离矩阵，当选中“从数据创建距离（Create distance from data）”选项时才可用。

距离（Distances）栏，此栏用于指定计算距离的方法。有如下两种方法。

数据为距离数据（Data are distances）选项：表示当前距离数据就是矩阵，可以直接用于分析。单击“形状（Shape）”按钮，则弹出如图 16-13 所示的“形状（Shape）”设置对话框，各选项功能如下所述。

- 对称正方形（Square Symmetric）：距离矩阵完全对称，行和列表示相同项目。
- 不对称正方形（Square Asymmetric）：距离矩阵为不对称方阵，行列表示相同项目。
- 矩形（Rectangular）：距离矩阵为完全不对称的矩阵形式，行和列表示不相同的项目。SPSS 把有序排列的数据当做矩形矩阵，如果其中含有多个矩形矩阵，则要设置每个矩阵的行数，其下的行数（Number of Rows）选项框用于指定单个矩阵行数，输入值应大于等于 4，且可以整除数据的所有行数。

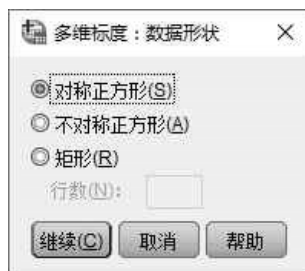


图 16-12 “多维标度 (Multidimensional Scaling)”对话框 图 16-13 “形状 (Shape) 设置”对话框

根据数据创建距离 (Create distances from data) 选项：表示用户需要自行选择相识矩阵的计算方法。当数据比较复杂、不可以直接用做距离矩阵时选择此项，表示从当前数据出发计算距离矩阵，单击“测量 (Measure)”按钮，则弹出如图 16-14 所示的“测量 (Measure)”对话框。

- 测量 (Measure) 栏：用于指定不相似度的测量方法。
- 转换值 (Transform Value) 栏：指定标准化转换的方法，这两部分的参数设置。
- 创建距离矩阵 (Create Distance Matrix) 栏：指定创建距离矩阵的方式。有两个选项“变量间 (Between Variables)”和“个案间 (Between Cases)”，分别表示计算配对变量之间的不相似距离矩阵和计算配对观测量之间的不相似性距离矩阵。

2. 模型设置

单击如图 16-12 所示的“模型 (Model)”按钮，则弹出如图 16-15 所示的“模型设置”对话框，各栏具体功能如下所述。



图 16-14 “测量 (Measure) 设置”对话框

图 16-15 “模型 (Model) 设置”对话框

(1) 测量级别 (Level of Measurement) 栏

此栏用于指定数据的测量尺度。

- 有序 (Ordinal): 表示有序测量尺度, 即分析数据是有序分类资料。其下“解除绑定已绑定的观察值 (Untie Tied Observations)”用于设置对节的处理方式。默认情况下, 对取值相同的评分赋予相同的秩。
- 区间 (Interval): 区间尺度。分析数据是由连续性变量或定量变量组成的。
- 比率 (Ratio): 比例尺度。分析数据是由比例形式的定量变量组成的。

(2) 条件性 (Conditionality) 栏

用于指定哪些比较是有意义的。

- 矩阵 (Matrix): 适用于只有一个距离矩阵, 或者每个距离矩阵仅代表单个受访者的情况。
- 行 (Row): 适用于距离矩阵为非对称或者矩形矩阵。
- 无约束 (Unconditional): 不受任何限制, 任意两个数据之间的比较都是有意义的。

(3) 维数 (Dimensions) 栏

用于指定尺度分析的维度, 系统默认为二维。“最小值 (Minimum)”和“最大值 (Maximum)”用于分别指定维数的最小值和最大值, 它们的输入值都需要为 1~6 的整数, 对指定范围内的每个维度分别进行分析。

(4) 标度模型 (Scaling Model) 栏

用于设置尺度模型的距离选项, 有两个选项。

- 欧氏距离 (Euclidean Distance)。
- 个别差异欧氏距离 (Individual differences Euclidean distance): 表示使用个体差异的欧氏距离矩阵进行分析, 它要求数据包含两个以上的距离矩阵。其下的“允许负的主题权重 (Allow Negative Subject Weights)”复选框表示允许权重变量为负值。

3. 选项设置

单击图 16-12 中的“选项 (Options)”按钮, 则弹出如图 16-16 所示的“输出设置”对话框, 单击“继续 (Continue)”按钮则返回原界面。

(1) 显示 (Display) 栏

此栏用于选择输出哪些图形和分析结果。

- 组图 (Group Plots): 多维标度分析图。
- 个别主体图 (Individual Subject Plots): 为每位受试者分别输出分析图形。
- 数据矩阵 (Data Matrix): 输出每位受试者的数据矩阵。
- 模型和选项摘要 (Model and Options Summary): 输出分析所有的数据、模型、算法等信息。

(2) 条件 (Criteria) 栏

此栏用于设置迭代收敛的依据。

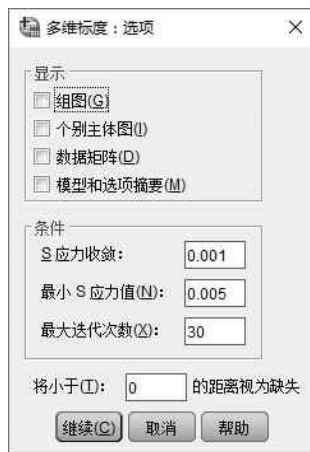


图 16-16 “输出 (Display) 设置”对话框

- S 应力收敛性 (S-stress Convergence): 指定 S-stress 的最小改变量, 默认值为 0.001。
- 最小 S 应力值 (Minimum s-stress value): 指定 S-stress 的最小值, 默认值为 0.005。
- 最大迭代次数 (Maximum Iterations): 指定最大的迭代次数。系统默认值为 30。

(3) 将小于“一个数”的距离视为缺失值 (Treat distances less than) 栏

此栏用于把距离小于某值的数据当做缺失值。

16.2.3 实例分析



结果文件

——附带光盘“PROGRAM\CH16\实例 16-2”文件夹



动画演示

——附带光盘“AVI\实例 16-2.avi”文件

本实例分析的数据集是 SPSS 自带的 kinship_dat.sav, 此数据集是 6 位被调查者对 15 位亲属的评分, 其中分数范围为 0~100 分, 分数越高说明关系越好。数据集 kinship_dat.sav 的格式如图 16-17 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|----|----------|----|----|------|---------------|---|----|---|----|----|----|
| 1 | aunt | 数字 | 8 | 2 | Aunt | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 2 | brother | 数字 | 8 | 2 | Brother | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 3 | cousin | 数字 | 8 | 2 | Cousin | 无 | 无 | 5 | 右 | 标度 | 输入 |
| 4 | daughter | 数字 | 8 | 2 | Daughter | 无 | 无 | 7 | 右 | 标度 | 输入 |
| 5 | father | 数字 | 8 | 2 | Father | 无 | 无 | 5 | 右 | 标度 | 输入 |
| 6 | gdaugh | 数字 | 8 | 2 | Granddaughter | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 7 | gfather | 数字 | 8 | 2 | Grandfather | 无 | 无 | 5 | 右 | 标度 | 输入 |
| 8 | gmother | 数字 | 8 | 2 | Grandmother | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 9 | gson | 数字 | 8 | 2 | Grandson | 无 | 无 | 5 | 右 | 标度 | 输入 |
| 10 | mother | 数字 | 8 | 2 | Mother | 无 | 无 | 6 | 右 | 标度 | 输入 |

图 16-17 数据格式

1. 参数设置

选择菜单“分析 (Analyze) 测量 (Scale) 多维标度 (Multidimensional Scaling)”, 则弹出“进行多维标度分析”对话框, 如图 16-18 所示。选中除变量 sourceid 以外的 15 个变量到“变量 (Variables)”选项栏中。

选中“选项 (Options)”按钮, 则弹出如图 16-19 所示对话框, 选中“组图 (Group Plots)”选项, 然后单击“继续 (Continue)”按钮返回主界面。

2. 结果分析

设置完成以后单击“确定”按钮进行分析, 如图 16-20 所示, 为迭代记录和相关性的输出。从图中可以看出迭代 5 次以后, S-stress (应力) 值的变化小于 0.001, 达到了收敛的标准。



图 16-18 “多维标度 (Multidimensional Scaling) 分析”对话框 图 16-19 “选项 (Options) 设置”对话框

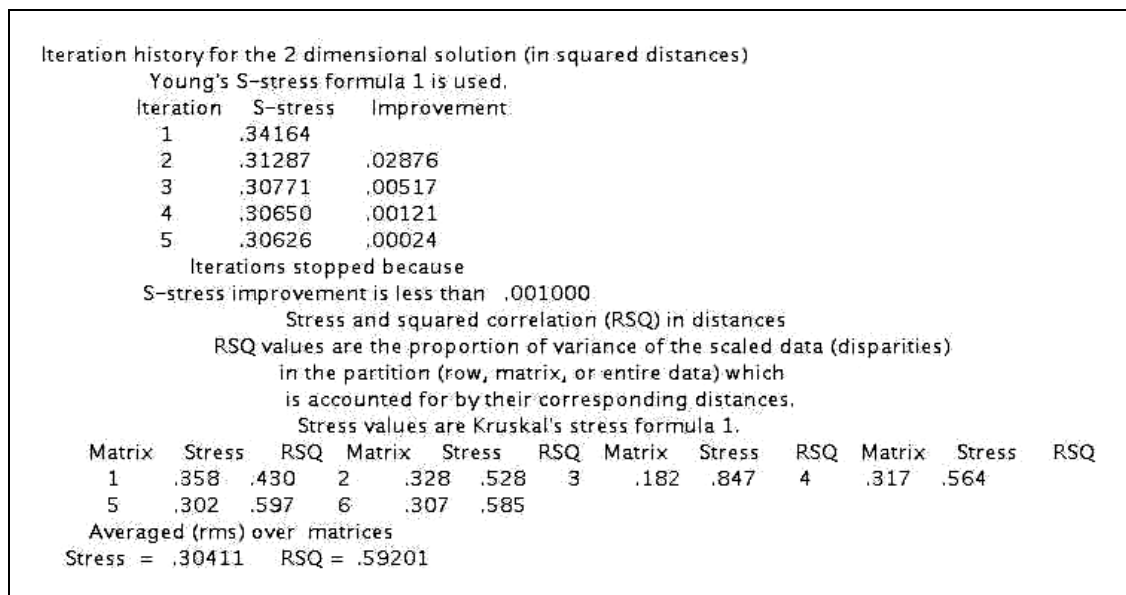


图 16-20 分析迭代距离和相信输出结果

图 16-21 给出的是二维导出构型表格，输出的是 15 个变量在二维空间中的坐标值，用于作多维标度分析图。

最后输出的是多维标度分析图形，图 16-22 所示为距离模型图形，它把反映变量之间相似程度的坐标在平面上排列出来，通过观察哪些散点比较接近，将变量进行分类，并寻找散点之间的相关性。图 16-23 是线性拟合图形，是欧氏距离对原始数据不一致程度的散点图，如果模型的拟合效果好，则所有散点应该分布在一条直线的周围，否则表示拟合效果不好。

| Stimulus Coordinates | | | | |
|----------------------|----------|---------|---------|--|
| Dimension | | | | |
| Stimulus | Stimulus | 1 | 2 | |
| Number | Name | | | |
| 1 | aunt | .6487 | 1.3888 | |
| 2 | brother | -1.0104 | -.8985 | |
| 3 | cousin | -.6467 | 1.7581 | |
| 4 | daughter | 1.2977 | -.3749 | |
| 5 | father | -.8394 | -1.0504 | |
| 6 | gdaugh | 1.3621 | -.0717 | |
| 7 | gfather | -1.0404 | -.9327 | |
| 8 | gmother | 1.3465 | -.2523 | |
| 9 | gson | -1.0398 | -.8261 | |
| 10 | mother | 1.2916 | -.3717 | |
| 11 | nephew | -1.2415 | .7089 | |
| 12 | niece | .6253 | 1.2928 | |
| 13 | sister | 1.2629 | -.1575 | |
| 14 | son | -.7936 | -.9857 | |
| 15 | uncle | -1.2230 | .7730 | |

图 16-21 二维导出构型表格

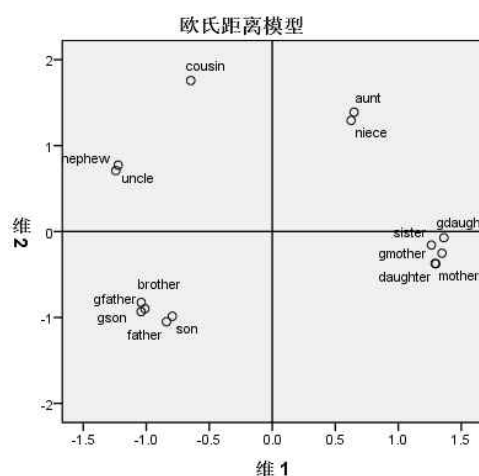


图 16-22 距离模型图形

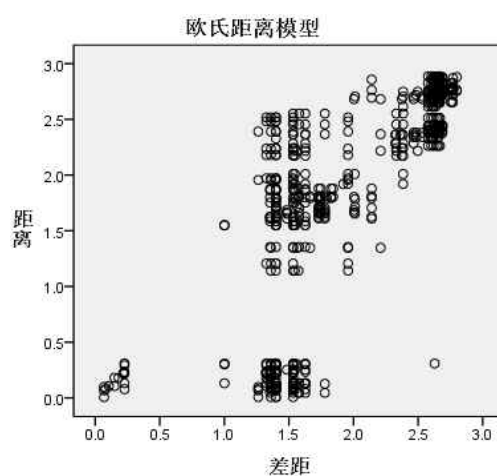
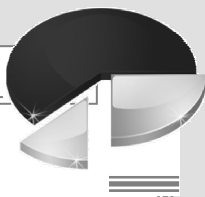


图 16-23 线性拟合散点图



第 17 章 生存分析

生存分析是一种以删失数据为研究对象的统计方法,近十年越来越受到人们的关注。1986 年美国国家科学院委员会提出的数学发展概况中,把生存分析列为六大发展方向之一,并作为数学与其他学科甚至社会科学互相渗透的一个重要例证。

生存分析是以人口寿命表为基础发展起来的,处理收集来的生存数据,生存数据包括生存时间,以及与其相关的因素。生存分析过去研究的主要领域是生物医学,而在工业、商业、社会科学等领域的应用也日渐扩大。

本章将利用 SPSS 软件进行生存分析。



本讲内容

- 生存分析基本概述
- Life Tables 过程
- Kaplan-Meier 分析
- Cox 模型回归分析

17.1 生存分析简介

生存分析的主要研究内容如下:描述生存过程,研究人群生存状态的规律;研究生存率曲线的变动趋势,它是人寿保险业的基础。生存过程影响因素分析及结局预测:了解哪些因素会影响生存过程;对生存结局加以预测,它在临床中的应用非常广泛。

17.1.1 生存分析的基本概念

首先了解一下本章中常用的术语。

1. 失效事件

也称“死亡”事件或失败事件,表示观察到随访对象出现了所规定的结局。失效事件的认定是生存分析的基石,必须绝对准确。失效事件应当由研究目的决定,并非一定是死亡(如研究灯泡寿命),而死亡也被并非一定是发生了失效事件(如肺癌患者死于其他疾病)。

2. 截尾值

终止随访不是由于失效事件发生,而是无法继续随访下去,常用符号“+”表示。生存但中途失访。包括拒绝访问、失去联系或中途退出试验。死于其他与研究无关的原因:如肺癌患者死于心肌梗塞、自杀或因车祸死亡,终止随访时间为死亡时间。随访截止:随访研究结束时观察对象仍存活。

3. 生存时间

生存时间(Survival Time)是指从某起点开始到被观测对象出现终点事件所经历的时间。如创业企业始创到创业失败等。由此可见,此处的“生存”是一个广义的概念。

生存时间常用下列三个函数来描述:生存函数、概率密度函数和危险率函数。

(1) 生存函数

生存函数又称累积生存率,记作 $S(t)$,指的是个体生存时间长于 t 的概率,即

$$S(t) = P\{\text{个体生存时间 } T > t\} = 1 - F(t)$$

式中, $F(t)$ 指个体的生存时间 T 的分布函数。假设生存率(Survival Rate)用 $S(t_k)$ 表示,指个体经历 t_k 个单位时间后仍存活的概率。若无删失数据,则

$$S(t_k) = P\{T \geq t_k\} = \frac{\text{过了 } t_k \text{ 时仍存活的个数}}{\text{观察开始时的总个数}}$$

其中 T 为个体的存活时间,但如果资料中含有删失数据,生存率的计算公式为

$$S(t_k) = P\{T \geq t_k\} = p_1, p_2, \dots, p_k$$

其中 p_1, p_2, \dots, p_k 表示不同时间段的生存概率,可以看出,这种情况下生存率是多个时间段生存率的累积,所以,又称累积生存概率。

(2) 概率密度函数

又称为密度函数,记作 $f(t)$,其函数表达式为

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P\{\text{个体在区间}(t, t+\Delta t)\text{中死亡}\}$$

(3) 危险率

危险率函数又称风险函数、危险率、死亡强度、条件死亡率,记作 $h(t)$,用于测量一定年龄的个体是否容易死亡。

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P\{\text{年龄是 } t \text{ 的个体在区间}(t, t+\Delta t)\text{中死亡}\}$$

风险函数的不同情况如下。

- 常数,如死于飞机失事。
- 下降,如急性损伤。
- 上升,如持续接触危险因素。
- 澡盆样式,即先上升,中间平稳,最后下降。如人的一生,也是如此(表示变化趋势)。

(4) 三者关系

生存函数、概率密度函数和危险率函数在数学上是等价的, 得出其中一个, 可以推导出另外两个, 其关系为

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1-F(t)}$$

4. 生存数据的删失

在生存分析中, 由于研究结束时, 某些个体可能还没有出现我们关心的事件, 那么, 这些个体的确切生存时间是不知道的, 这就导致生存数据存在删失情况。正是由于生存分析存在着删失数据, 使得传统的分析方法可能在分析时出现偏差。生存数据一般包括以下几个方面。

完全数据: 是指被观测对象的观测数据完全落于观察起点至终点间。

删失数据 (Censored Data): 包括左删失和右删失。一般常见的是右删失, 它是指在出现终点事件前, 被观测对象的观测过程由于各种原因而终止了。右删失主要有以下三种情况。

- 失访: 指失去联系, 如创业企业搬迁而失去联系。
- 退出: 指退出研究, 如因其他原因、临时改变方案而中途退出研究。
- 终止: 指研究时限已到而终止观察。

17.1.2 生存资料的特点

生存期不同于一般指标的两个特点。

1. 有截尾数据 (Censored Data)

随访中未能知道病人的确切生存时间, 只知道病人的生存时间大于某时间。

病人失访或因其他原因而死亡——失访。

到了研究的终止期病人尚未死亡——终访。

截尾数据可记为 T^+ , 如 4^+ = 生存时间大于 4 年。虽然截尾数据提供的信息是不完全的, 但不能删去, 因为这不仅损失了资料, 而且会造成偏差。

2. 生存期的资料一般不服从正态分布

由于上述原因, 常用的统计方法不适用, 而要用特殊的统计方法。生存分析是指对于生存期这一指标进行分析的一系列特殊的统计方法。

生存时间不一定专用于死与活的情况, 生存时间 (存活时间) 可定义为从某种起始事件到达某终点事件所经历的时间跨度。例如, 急性白血病病人从治疗开始到复发为止之间的缓解期; 冠心病病人在两次发作之间的时间间隔。在流行病学研究中, 从开始接触危险因素到发病所经历的时间等都可作为生存时间用做生存分析。

有时还收集一些有关因素 (称为自变量或协变量), 以分析这些协变量是否对生存时间有影响, 影响的大小, 是缩短或延长生存时间。这可以通过 Cox 回归进行分析, 因此, Cox 回归可看做带有协变量的生存分析。

17.1.3 生存分析方法

目前,生存分析的基本方法大体上有寿命表分析、Kaplan-Meier 分析、Cox 回归模型等。

1. 寿命表分析

寿命表法(Life-Table Method, LT 法)是通过计算落入时间区间 $[t_{k-1}, t_k]$ 内的失效和删失的观察个数来估计该区间上的死亡概率,然后用该区间及其之前各区间上的生存概率之积来估计 $S(t_k)$ 。

2. Kaplan-Meier 分析

又称乘积极限法(Product-Limit Method, PL 法),是 Kaplan 与 Meier 于 1958 年提出的。它可以运用删失数据建立时间-事件模型,根据每一个事件发生时间点的条件概率的估计和事件相应的概率等信息来估计每一个时点的生存率。

3. Cox 回归分析

Cox 回归分析是一种存在删失数据情况下拟合时间-时间模型的一种方法。Cox 回归模型中可包含预测变量(协变量),当众多的危险因素对生存时间有影响时,应关心其中哪些危险因素对生存时间有重要的影响,也就是确认重要的预后因素,通过建立生存时间随危险因素变化的回归模型,确定这些对生存时间有影响的预后因素,并根据危险因素在模型中的影响对生存率进行预测。具体情况在第 17.4 节详述。

生存分析中应用最多的多因素分析方法是 Cox 模型。而 Cox 模型是一种半参数模型,它是在假设不同个体的死亡风险在所有时间上都保持一个恒定比例的条件下提出的,但是实际情况多不能满足这个条件;寿命表法适用于数据按区间分组或者大样本,以及无法准确得知研究结果出现时间的情况,Kaplan-Meier 法主要用于一个个的个体数据,而非分组数据,但是当每个分组区间只包含一个数据时,Kaplan-Meier 方法可看做是寿命表分析的特例。

总之,生存分析一般可以分为参数、非参数、半参数三类。

参数法:生存时间的分布符合某一特定类型,如对数正态分布、Weibull 分布、指数分布、Gamma 分布等,则可用特定的分布函数分析,这称为参数法。

非参数法:用 Kaplan-meier 法或寿命表法求生存率,作生存曲线;用 Logrank 检验或 Breslow 检验比较两组或几组生存率差异有无统计学意义。

半参数法:Cox 比例风险模型。

17.1.4 SPSS 中的生存分析过程

SPSS 中提供了很全面的生存分析处理过程,主要有如下四个分析过程。

- 寿命表(Life Tables):寿命表适用于分组生存资料分析过程,可以求出不同组段的生存率。
- Kaplan-Meier:适用于样本比较小的观测,它不能给出特定时刻的生存率。
- Cox 回归(Cox Regression):用于进行拟合 Cox 比例风险模型。

下面将对比较常用的寿命表 (Life Tables) 过程、Kaplan-Meier 过程, 以及 Cox 回归过程进行介绍, 并以实例来进行分析。

17.2 寿命表 (Life Tables) 过程

17.2.1 寿命表分析过程的参数设置

选择菜单“分析 (Analyze) 生存分析 (Survival) 寿命表 (Life Tables)”命令, 则系统执行生存分析过程, 弹出的对话框如图 17-1 所示。

图 17-1 中的左边是待分析变量列表, 下面就各个选项框的功能进行说明。

1. 时间 (Time) 选项栏

此栏用于选入代表生存时间的变量, 其下面的“显示时间间隔 (Display Time Intervals)”子设置栏用于指定寿命表中生存时间的范围及其组距, “到 (through)”后输入指定的生存时间的上限; “按 (by)”后输入指定生存时间的组距。

2. 状态 (Status) 选项栏

此栏用于选入定义事件是否发生的生存状态变量, 选入变量后会自动激活“定义事件 (Define Event)”按钮, 单击此按钮, 则弹出如图 17-2 所示对话框, 此对话框中各选项功能如下所述。

- 单值 (Single Value): 当生存状态为二元变量时, 选中此项, 并在后面输入指定状态变量的代表事件发生的取值即可。
- 值的范围 (Range of Value): 当生存状态为多分类变量时, 选中此项, 并在“到 (through)”前输入框指定取值范围的起始值, 在“到 (through)”后输入指定取值范围的终止值。

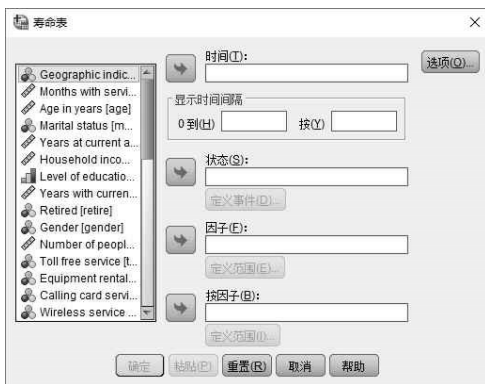


图 17-1 “寿命表 (Life Tables)”对话框



图 17-2 “定义事件 (Define Event)”对话框

3. 因子 (Factor) 选项栏

此栏用于选入第一个因素变量, 用于分组。选入变量后自动激活“定义范围 (Define Range)”按钮, 单击此按钮, 则弹出如图 17-3 所示对话框, 此对话框中各选项功能如下所述。

- 最小值 (Minimum): 输入取值范围的最小值。
- 最大值 (Maximum): 输入取值范围的最大值。

4. 按因子 (By Factor) 选项栏

此选项栏用于选入第二个因素变量, 设置同因子 (Factor) 选项栏。

5. 选项 (Options) 设置

单击图 17-1 中的“选项 (Options)”按钮, 则弹出如图 17-4 所示的对话框, 此对话框主要设置有关图形、表格的输出, 具体功能如下。

寿命表 (Life Tables): 在结果里输出寿命表。

图 (Plot) 栏: 此栏用于设置输出图形的类型。

- 生存分析 (Survival): 累积生存函数曲线。
- 生存函数对数 (Log Survival): 对数累积生存曲线。
- 风险 (Hazard): 累积风险函数散点图。
- 密度 (Density): 密度函数散点图。
- 一减去生存分析函数 (One Minus Survival): 生存函数被 1 减去后的曲线图。

比较第一个因子的级别 (Compare Levels of First Factor) 栏: 此栏设置对第一个因素不同取值水平的比较方法。

- 无 (None): 不做比较, 此项为系统默认。
- 总体 (Overall): 整体比较, 其检验的零假设为各分组的生存曲线全部相同, 相当于方差分析中的总体比较。
- 成对 (Pairwise): 相当于方差分析中的两两比较。



图 17-3 “定义因子范围 (Define Range) 量设置”对话框



图 17-4 “选项”对话框

17.2.2 实例分析



结果文件

——附带光盘“PROGRAM\CH17\实例 17-1”文件夹



动画演示

——附带光盘“AVI\实例 17-1.avi”文件

企业为了有效地保留客户，有必要建立有效的客户流失模型，快速、准确了解客户的流失情况。如每个客户流失的概率有多大，某个客户流失的原因是什么，谁是潜在的流失客户等，并据此有针对性地制定营销策略，采取行动让客户满意，留住客户，将最好的客户留住更长的时间，以提升客户存在期的价值，最终达到减少客户流失的目的。

客户流失数据集本身存在如下的特殊性。

第一，客户流失分析所用的数据集包含已经流失的用户（流失用户）和还在使用的用户（在用用户）的数据，由于在用用户随时都可能流失，使用上述数据挖掘工具就会碰到问题，即数据集中可能有很多用户，收集分析数据时是在用用户，但结果出来时已经发生了流失，这样就会给挖掘工具一个错误的指导信息，挖掘工具根据这个指导信息建立的模型的可靠性就很值得怀疑。

第二，截至分析的时候，很多客户还是在用用户，还没有发生流失，故不知道他们流失前使用时间的确切值，只知道肯定大于等于某个数，由于某种原因被截断了。如果此时使用 Logistic 回归等挖掘工具进行分析，由于这些工具是针对没有截断的完全数据进行分析的，截断值并不是真实值，所以得到的分析结果将偏低失真。

对于这种包含生存时间不能准确观测到的对象，既不能简单地弃之，又不能充分信任的信息，生存分析方法提供了很好的解决方法。

本实例使用 SPSS 自带的数据集 telco.sav 来进行寿命表的分析。此数据集为某电信公司的客户数据集，其 SPSS 数据格式如图 17-5 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|----|---------|----|----|------|--------------------|------------------|----|----|----|----|----|
| 1 | region | 数字 | 4 | 0 | Geographic indi... | {1, Zone 1}... | 无 | 6 | 右 | 名义 | 输入 |
| 2 | tenure | 数字 | 4 | 0 | Months with se... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 3 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 4 | marital | 数字 | 4 | 0 | Marital status | {0, Unmarrie... | 无 | 7 | 右 | 名义 | 输入 |
| 5 | address | 数字 | 4 | 0 | Years at curren... | 无 | 无 | 7 | 右 | 标度 | 输入 |
| 6 | income | 数字 | 8 | 2 | Household inco... | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 7 | ed | 数字 | 4 | 0 | Level of education | {1, Did not c... | 无 | 6 | 右 | 有序 | 输入 |
| 8 | employ | 数字 | 4 | 0 | Years with curr... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 9 | retire | 数字 | 8 | 2 | Retired | {00, No}... | 无 | 10 | 右 | 名义 | 输入 |
| 10 | gender | 数字 | 4 | 0 | Gender | {0, Male}... | 无 | 6 | 右 | 名义 | 输入 |

图 17-5 数据集 telco.sav 的格式

1. 参数设置

选择菜单“分析（Analyze）生存分析（Survival）寿命表（Life Tables）”命令，则弹出如图 17-6 所示对话框，此对话框用于寿命表分析过程中的参数设置。

如图 17-6 所示，选入变量 Months with service 到“时间（Time）”变量框中，其下的“显示时间间隔（Display Time Intervals）”选项栏中，设置到（through）为 60，步长（by）为 3。选中变量 Churn 到“状态（Status）”变量框中。选中变量 custcat 到“因子”变量框。

选入变量后，再进行变量状态（Status）和变量因子（Factor）的参数设置，单击“状态（Status）”选项栏下的“定义事件（Define Event）”按钮，则弹出如图 17-7 所示的对话框。在“单值（Single Value）”选项栏中填入 1。然后单击“继续（Continue）”按钮返回主界面。

单击“因子（Factor）”选项栏下的“定义范围（Define Range）”按钮，则弹出如图 17-8

所示的对话框。在“最小值 (Minimun)”选项栏中填入 1, 在“最大值 (Maximun)”选项栏中填入 4, 然后单击“继续 (Continue)”按钮返回主界面。

单击图 17-6 中的“选项 (Options)”按钮, 弹出如图 17-9 所示对话框, 对话框中的各种设置从图 17-9 中可以看出。在“图 (Plot)”选项栏中选中“生存分析 (Survival)”选项来绘制寿命表的绘图, 在“比较第一个因子的级别 (Compare Levels of First Factor)”选项栏中选中“成对 (Pairwise)”选项, 然后单击“继续 (Continue)”按钮返回主界面。



图 17-6 “寿命表”对话框



图 17-7 “寿命表：为状态变量定义事件”对话框



图 17-8 “有效表格：定义因子范围”对话框



图 17-9 “寿命表：选项”对话框

设置完上述参数以后, 则单击“寿命表”对话框 (Life Tables Dialog Box) 界面中的“确定 (OK)”按钮, 然后系统进行分析。

2. 结果分析

单击“确定 (OK)”按钮以后, 则输出分析结果, 首先是寿命表的输出, 显示的是基本的统计信息, 由于此表数据较大, 在此不再给出。

图 17-10 给出的是客户流失累积生存函数图形。此图形可以让用户比较直观地观测寿命表的输出结果。从图 17-10 中的输出结果可以看出, 基本服务 (Basic Service) 的客户累积生存函数下降最快, 而网络服务 (E-service) 比附加服务 (Plus Service) 的客户累积生存函数下降得快。

然后输出的是整体检验结果, 如图 17-11 所示, 从结果可以看出显著性值为 0.000, 表

明这四种客户生存曲线是显著不同的。

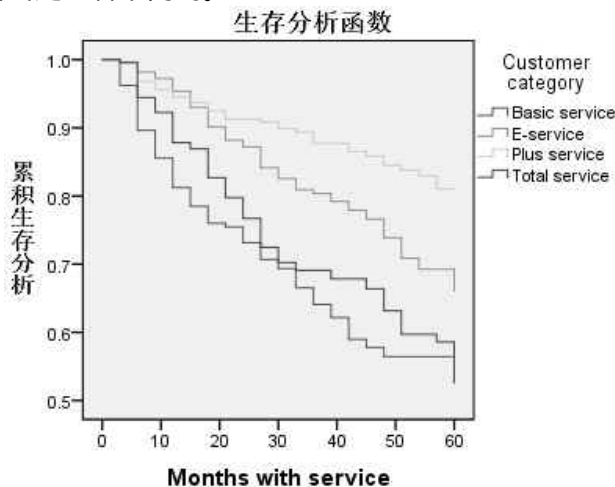


图 17-10 客户流失累积生存函数图

| 总体比较 ^a | | |
|-------------------|-----|------|
| 威尔科克森 (吉亨)统计 | 自由度 | 显著性 |
| 49.179 | 3 | .000 |
| a. 执行的是精确比较。 | | |

图 17-11 整体检验结果

图 17-12 给出的是成对比较的结果，此表给出了更加详细的结果。表明了基本服务的客户和上网、附加服务的客户生存曲线之间的差异是显著性的，但是基本服务和所有服务 (Total Service)，以及上网服务和附加服务的客户之间的生存曲线差异是不显著的。

| 成对比较 ^a | | | | |
|-------------------|-------------|-----------------|-----|------|
| (I) custcat | (J) custcat | 威尔科克森 (吉亨)统计 | 自由度 | 显著性 |
| 1 | 2 | 18.640 | 1 | .000 |
| | 3 | 37.154 | 1 | .000 |
| | 4 | 2.949 | 1 | .086 |
| 2 | 1 | 18.640 | 1 | .000 |
| | 3 | 5.515 | 1 | .019 |
| | 4 | 9.222 | 1 | .002 |
| 3 | 1 | 37.154 | 1 | .000 |
| | 2 | 5.515 | 1 | .019 |
| | 4 | 27.229 | 1 | .000 |
| 4 | 1 | 2.949 | 1 | .086 |
| | 2 | 9.222 | 1 | .002 |
| | 3 | 27.229 | 1 | .000 |
| a. 执行的是精确比较。 | | | | |

图 17-12 成对比较结果

17.3 Kaplan-Meier 分析

17.3.1 Kaplan-Meier 分析过程的参数设置

选择菜单“分析 (Analyze) 生存分析 (Survival) Kaplan-Meier”命令, 则系统执行 Kaplan-Meier 生存分析过程, 弹出的对话框如图 17-13 所示。

1. 时间 (Time) 选项

此选项用于选中生效时间变量。

2. 状态 (Status) 选项

此选项用于选入生存状态变量。选入变量后, 系统会自动激活“定义事件 (Define Event)”按钮, 单击此按钮, 则会弹出图 17-14 所示对话框。各选项栏功能如下所述。



图 17-13 “Kaplan-Meier 过程”对话框



图 17-14 “定义事件设置”对话框

- 单值 (Single Value): 当生存状态为二元变量时, 选中此项, 并在后面的输入框中指定状态变量的代表事件发生的取值即可。
- 值的范围 (Range of Value): 当生存状态为多分类变量时, 选中此项, 并在“到 (through)”前输入框指定取值范围的起始值, 在“到 (through)”后输入指定取值范围的终止值。
- 值的列表 (List of Values) 栏: 在其后输入某个数字, 单击“添加 (Add)”按钮将加入下面的列表中, 如此重复可以指定代表事件发生的多个不同的值; 如果需要更改已经填入的值, 则可以在列表选中, 然后在“值的列表 (List of Values)”输入框中进行编辑, 最后单击更改“(Change)”按钮即可, 单击“除去 (Remove)”按钮则可以直接删除所选中的值。

3. 因子 (Factor) 选项

此选项用于选入因素变量。

4. 层 (Strata) 选项

此选项用于选入分层因素。

5. 个案标注依据 (Label Cases by) 选项

此选项用于选入观测的标签变量。

6. 比较因子 (Compare Factor) 设置

此选项是因素取值水平的比较设置，单击“比较因子 (Compare Factor)”按钮，则弹出如图 17-15 所示对话框，此对话框可以设置对因素变量取值水平的比较方法。各个选项功能如下所述。

(1) 检验统计量 (Test Statistics) 栏

此栏用于选择具体的检验统计量。

- 秩的对数 (Log Rank)：检验各组生存率曲线分布是否相同，且各时刻权重一样。
- 布雷斯洛：检验各组生存率曲线的分布是否相同，并以各时刻的观察例数为权重。
- 塔罗内-韦尔：检验各组生存率曲线的分布是否相同，以各个时刻观察例数的平方根为权重。

(2) 因子级别的线性趋势 (Linear Trend for Factor Levels) 选项

此栏用于指定分组因素各水平之间的线性趋势检验。只有当分组因素是有序变量时，线性趋势检验才有实际意义，此种情况之下，SPSS 假设各水平之间的效应是等距的。

(3) 图 17-15 最后的一组单选框用来指定进行总体比较还是两两比较，以及分层变量的处理方式，各选项含义如下。

- 在层之间汇聚 (Pooled over Strata)：对各因素变量取值水平下的生存曲线作整体比较，此为默认选项。
- 针对每个层 (For each Stratum)：按照分层变量的不同取值，对每一层分别进行因素变量各取值水平间的整体比较，如果没有指定分层变量，则不会输出。
- 在层之间成对比较 (Pairwise over Strata)：进行因素变量各水平之间的两两比较。对线性趋势检验无效。
- 针对每个层成对比较 (Pairwise for each Stratum)：按照分层变量的不同取值，对每一层分别进行因素变量各取值水平间的两两比较。对线性趋势检验无效。

7. 保存 (Save) 设置

单击图 17-13 中的“保存 (Save)”按钮，则弹出如图 17-16 所示的对话框，此选项框用于设置保存的选项。



图 17-15 “比较因子水平”对话框



图 17-16 “保存新变量”对话框

- 生存分析 (Survival): 累积生存率估计值。
- 生存分析的标准误差 (Standard error of survival): 累积生存率估计值的标准差。
- 风险 (Hazard): 累积风险函数的估计值。
- 累积事件 (Cumulative Events): 终结时间的累积频数。

8. 选项 (Options) 设置

单击图 17-13 中的“选项 (Options)”按钮, 则弹出如图 17-17 所示的对话框, 此选项框用于选择要输出的分析结果。

(1) 统计量 (Statistics) 栏

此栏设置要输出的统计量, 可以选择的内容有以下几个方面。

- 生存分析表 (Survival Table (s)):
- 平均值和中位数生存分析函数 (Mean and Median Survival): 平均生存时间和中位生存时间, 以及各自的标准误差和置信区间。
- 四分位数 (Quartiles): 输出生存时间的四分位数。

(2) 图 (Plots) 栏

此栏用于设置要输出的统计图形, 可选内容如下。

- 生存分析函数 (Survival): 累积生存函数曲线。
- 一减去生存函数 (One Minus Survival): 生存函数减去 1 所得的曲线。
- 风险 (Hazard): 累积风险函数的散点图。
- 生存分析函数的对数 (Log Survival): 对数累积生存函数曲线。

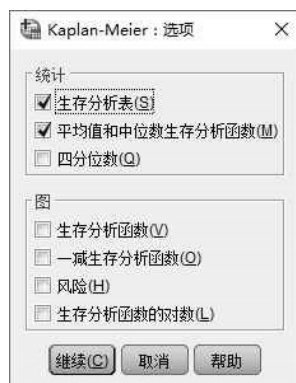


图 17-17 “选项”对话框

17.3.2 实例分析



结果文件

——附带光盘“PROGRAM\CH17\实例 17-2”文件夹



动画演示

——附带光盘“AVI\实例 17-2.avi”文件

本实例还是选择 SPSS 自带的数据集 pain_medication.sav, 此数据集是关于药物调查数据, 以利于新药物的开发。数据集 pain_medication.sav 的数据格式如图 17-18 所示。下面对此数据集进行寿命表的分析操作。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|-----------|----|----|------|----------------|----------------|----|----|----|----|----|
| 1 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 2 | gender | 数字 | 4 | 0 | Gender | {0, Male}... | 无 | 6 | 右 | 名义 | 输入 |
| 3 | health | 数字 | 4 | 0 | General health | {1, Poor}... | 无 | 6 | 右 | 有序 | 输入 |
| 4 | treatment | 数字 | 4 | 0 | Treatment | {0, New dru... | 无 | 9 | 右 | 名义 | 输入 |
| 5 | dosage | 数字 | 4 | 0 | Dosage | {0, Low}... | 无 | 6 | 右 | 有序 | 输入 |
| 6 | status | 数字 | 4 | 0 | Effect status | {0, Censore... | 无 | 6 | 右 | 名义 | 输入 |
| 7 | time | 数字 | 8 | 2 | Time to effect | 无 | 无 | 10 | 右 | 标度 | 输入 |

图 17-18 数据集 pain_medication.sav 的数据格式

1. 参数设置

选择菜单“分析 (Analyze) 生存分析 (Survival) Kaplan-Meier”命令，则系统执行 Kaplan-Meier 生存分析过程，弹出的对话框如图 17-19 所示。选择变量 Time to effect 到“时间 (Time)”变量框之中，选择变量 Effect status 到“状态 (Status)”变量框之中。

选入变量之后单击“定义事件 (Define Event)”按钮，弹出如图 17-20 所示对话框，进行事件的设置操作。在“单值 (Single Value)”变量框中输入 1，然后单击“继续 (Continue)”按钮返回主界面。



图 17-19 “Kaplan-Meier”对话框



图 17-20 “定义状态变量事件”对话框

选择变量 Treatment 到“因子 (Factor)”变量框中，则在进行“比较因子 (Compare Factor)”选项框的设置。单击图 17-19 中的“比较因子 (Compare Factor)”按钮，弹出如图 17-21 所示对话框，各种选项设置如图 17-21 所示，然后单击“继续 (Continue)”按钮返回主界面。

单击图 17-19 中的“选项 (Options)”按钮，则弹出“选项”对话框。各种选项设置如图 17-22 中所示。单击“继续 (Continue)”按钮返回主界面。

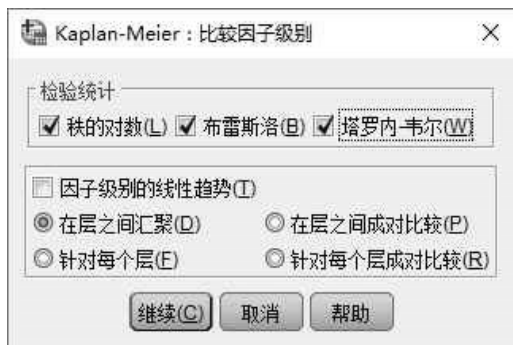


图 17-21 “比较因子水平”对话框

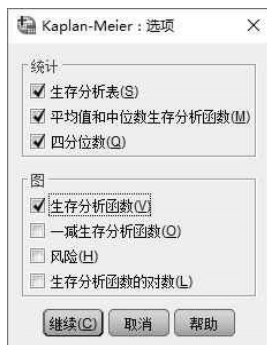


图 17-22 “选项”对话框

2. 结果分析

设置好各种参数以后，单击 Kaplan-Meier 对话框主界面中的“确定 (OK)”按钮进行分析，结果如下。首先是个体案例的分析结果，给出了样本数据的简要统计信息，如图 17-23 所示。

同寿命表分析结果一样，Kaplan-Meier 生存分析也给出寿命表分析结果。由于数据很多在此只给出部分内容，如图 17-24 所示。

| 个案处理摘要 | | | | |
|---------------|-----|-----|-----|-------|
| Treatment | 总数 | 事件数 | 检删后 | |
| | | | 个案数 | 百分比 |
| New drug | 104 | 79 | 25 | 24.0% |
| Existing drug | 96 | 74 | 22 | 22.9% |
| 总体 | 200 | 153 | 47 | 23.5% |

图 17-23 个体案例的分析结果

| 生存分析表 | | | | | | | |
|-----------|---|-------|--------------|------------|------|-------|-------|
| Treatment | | 时间 | 状态 | 当前累计生存分析比例 | | 累积事件数 | 其余个案数 |
| | | | | 估算 | 标准误差 | | |
| New drug | 1 | .600 | Taken effect | . | . | 1 | 103 |
| | 2 | .600 | Taken effect | .981 | .013 | 2 | 102 |
| | 3 | .700 | Taken effect | .971 | .016 | 3 | 101 |
| | 4 | .800 | Taken effect | .962 | .019 | 4 | 100 |
| | 5 | .900 | Taken effect | .952 | .021 | 5 | 99 |
| | 6 | 1.100 | Taken effect | . | . | 6 | 98 |
| | 7 | 1.100 | Taken effect | .933 | .025 | 7 | 97 |

图 17-24 寿命表分析结果

图 17-25 给出的是寿命表的均值、中位数，以及其他统计信息，图 17-26 给出的是寿命表的百分位数。

| 生存分析时间的平均值和中位数 | | | | | | | | |
|----------------|------------------|------|----------|-------|-------|-------|----------|-------|
| Treatment | 平均值 ^a | | | | 中位数 | | | |
| | 估算 | 标准误差 | 95% 置信区间 | | 估算 | 标准误差 | 95% 置信区间 | |
| | | | 下限 | 上限 | | | 下限 | 上限 |
| New drug | 4.867 | .360 | 4.162 | 5.572 | 3.700 | .292 | 3.128 | 4.272 |
| Existing drug | 5.185 | .350 | 4.499 | 5.871 | 4.100 | 1.131 | 1.884 | 6.316 |
| 总体 | 5.014 | .252 | 4.520 | 5.507 | 3.900 | .272 | 3.367 | 4.433 |

a. 如果已对生存分析时间进行检删，那么估算将限于最大生存分析时间。

图 17-25 寿命表统计信息

| 百分位数 | | | | | | |
|---------------|-------|------|-------|-------|-------|------|
| Treatment | 25.0% | | 50.0% | | 75.0% | |
| | 估算 | 标准误差 | 估算 | 标准误差 | 估算 | 标准误差 |
| New drug | 7.100 | .509 | 3.700 | .292 | 1.900 | .226 |
| Existing drug | 7.700 | .648 | 4.100 | 1.131 | 2.400 | .247 |
| 总体 | 7.300 | .371 | 3.900 | .272 | 2.100 | .196 |

图 17-26 寿命表的百分位数

最后输出的是整体比较和生命函数图形，如图 17-27、图 17-28 所示的输出结果。如图 17-27 所示的检验结果可以看到显著性的数值都很大，所以，在显著性的水平上没有显著性的差异。如图 17-28 所示的生存函数图形显示了新药的生存函数多位于旧药生存函数的下面，所以，新药的生效时间比旧药要好，但是从统计学上来看这种差异并不显著。

| 总体比较 | | | |
|-----------------------------------|------|-----|------|
| | 卡方 | 自由度 | 显著性 |
| Log Rank (Mantel-Cox) | .379 | 1 | .538 |
| Breslow (Generalized Wilcoxon) | .748 | 1 | .387 |
| Tarone-Ware | .705 | 1 | .401 |
| 针对 Treatment 的不同级别进行的生存分析分布等同性检验。 | | | |

图 17-27 整体比较检验结果

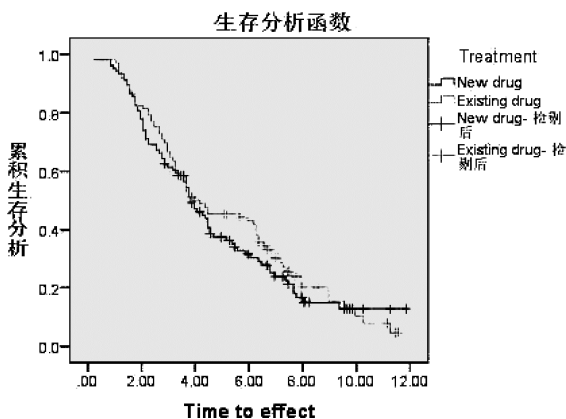


图 17-28 生命函数图形

17.4 Cox 模型回归分析

Cox 回归是一种允许资料有“删失（或截尾）”数据存在的，可以同时分析众多因素对生存时间影响的多变量生存分析方法，是一种半参数方法。

Cox 回归用于研究各种因素（称为协变量，或伴随变量等）对于生存期长短的关系，进行多因素分析。首先是 Cox 回归模型。

17.4.1 Cox 回归模型

设有 n 名病人 ($i = 1, 2, \dots, n$)，第 i 名病人的生存时间为 t_i ，同时该病人具有一组伴随变量 $x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}$ ，则模型为

$$\ln h(t, X) = \ln h_0(t) + (\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)$$

$$h(t, X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)$$

$$\ln \frac{h(t, X)}{h_0(t)} = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

$$h(t, X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)$$

其中

- $h(t, X)$ 为在时间 t 处与 X （协变量）有关的风险函数（Hazard Function）。
- β 为回归系数（最大似然估计值记为 b ）。

- $h_0(t)$ 为基准 (Baseline) 风险函数, 是与时间有关的任意函数, 函数形式无任何限定。
- $\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ 称为预后指数 (Prognostic Index)。

每一病人死亡风险的比例系数为

$$\frac{h(t)}{h_0(t)} = \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

模型中有参数 β , 但基准风险函数没有定义, 故又称为半参数模型。

风险——指瞬间风险 (Instantaneous Hazard), 或死亡力 (force of mortality), 用 $h(t)$ 表示, 是在时间点 t 尚存个体在短暂时期 Δt 内发生死亡的危险程度。即指生存到时间 t 的病人, 从 t 到 $t + \Delta t$ 这一非常小时间区间内的瞬间死亡概率。

Kaplan-Meier 法计算的死亡概率 q_i 就是 $h(t)$ 的估计值。

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(\text{在 } (t, t + \Delta t) \text{ 瞬间死亡} | \text{在 } t \text{ 时刻尚存者})}{\Delta t}$$

英国生物统计学家 D.R.Cox 于 1972 年通过条件死亡概率建立偏似然函数 L_p , 使对数似然函数 $\log L_p$ 最大, 通过最大似然法的 Newton-Raphson 迭代得到参数 $\beta_1, \beta_2, \cdots, \beta_p$ 的估计值 b_1, b_2, \cdots, b_p , 则

$$\begin{aligned} L_p &= \prod_{i=1}^{d(\text{非删失时点数})} q_i \\ &= \prod_{i=1}^{d(\text{非删失时点数})} \frac{h_0(t_i) \exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip})}{\sum_{j \in R_i} h_0(t_i) \exp(\beta_1 X_{j1} + \cdots + \beta_p X_{jp})} \\ &= \prod_{i=1}^{d(\text{非删失时点数})} \frac{\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip})}{\sum_{j \in R_i} \exp(\beta_1 X_{j1} + \cdots + \beta_p X_{jp})} \end{aligned}$$

分母中 $j \in R_i$ 表示在 t_i 时刻的所有个体 (包括删失个体) 风险之和, 分子只反映观察到的死亡风险。只有非删失 (死亡) 个体才有偏似然函数。引入 $\delta_i = \begin{cases} 1, & \text{第 } i \text{ 个体死亡} \\ 0, & \text{第 } i \text{ 个体删失} \end{cases}$, 则 L_p

可以写成

$$L_p = \prod_{i=1}^n \left(\frac{\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip})}{\sum_{j \in R_i} \exp(\beta_1 X_{j1} + \cdots + \beta_p X_{jp})} \right)^{\delta_i}$$

令 $l(\beta) = \ln L_p$, 所以

$$l(\beta) = \ln L_p = \sum_{i=1}^d (\beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - \sum_{i=1}^d \ln \left\{ \sum_{j \in R_i} (\beta_1 x_{j1} + \cdots + \beta_p x_{jp}) \right\}$$

然后, 再令 $\frac{dl(\beta)}{d\beta} = 0$, 即可求解回归参数。

回归系数实际上是偏回归系数, 其意义与多元线性回归模型或 Logistic 回归模型中的

偏回归系数的意义相似。表示控制其他因素条件下，各个因素对回归方程的独立贡献。

观察值经过标准化变换后所求得回归系数称为标准偏回归系数 b'_j 。 $b'_j = S_j b_j$ ，是相对值用于比较自变量对于模型的贡献。

风险如下：

$$\text{风险 (Risk)} = \frac{h(t)}{h_0(t)} = \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

如 X_1 改变 1 个单位，即 $X_1 = a$ 改变到 $X_1 = a + 1$ 时，风险比 (Risk Ratio，记为 RR_t) 为

$$\text{风险比} = RR = \frac{\text{风险 2}}{\text{风险 1}} = \frac{\exp[b_1(a+1) + b_2 X_2 + \cdots + b_p X_p]}{\exp[b_1(a) + b_2 X_2 + \cdots + b_p X_p]} = \exp b_1$$

故回归系数 b_j 反映了其他自变量固定不变的情况下， X_j 改变 1 个单位， X_j 所引起的危险比改变量为 $\exp(b_j)$ 。

基准生存函数为

$$S_0(t_i) = \prod_{j=1}^i p_j = \prod_{j=1}^i (1 - q_j) = \prod_{j=1}^i \left(1 - \frac{d_j}{\sum_{l \in R_l} \exp(b_1 X_{l1} + \cdots + b_p X_{lp})} \right)$$

生存函数为

$$S(t_i) = S_0(t_i) \exp(b_1 X_1 + b_2 X_2 + \cdots + b_p X_p)$$

式中， $\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ 便是风险指数 (Hazard Index)，记为 HI，为了应用上的方便，通常用其标准化的估计值，即

$$HI = b'_1 X'_1 + b'_2 X'_2 + \cdots + b'_p X'_p$$


式中， X'_1, X'_2, \cdots, X'_p 为标准化自变量； b'_1, b'_2, \cdots, b'_p 是 Cox 模型标准化回归系数。因风险可决定预后，故风险指数又称预后指数 (Prognostic Index)，或预后得分 (Prognostic Score)。

- HI = 0 代表风险处于平均水平。
- HI < 0 表示风险低于平均水平。
- HI > 0 表示风险高于平均水平。

17.4.2 Cox 模型分析过程的参数设置

选择菜单“分析 (Analyze) 生存分析 (Survival) Cox 回归 (Cox Regression)”命令，则系统自动执行 Cox 回归模型分析，弹出如图 17-29 所示对话框。

1. 分析变量设置

如图 17-29 所示的左边是待分析变量列表，选中变量，然后单击按钮  即可选入对应的变量框之中。各变量框的含义如下所述。

- 时间 (Time)：选入时间变量。

- 状态 (Status) : 用于选入生存状态变量。选入变量后, “定义事件” 按钮会被激活, 单击此按钮, 则弹出如图 17-30 所示对话框, 此对话框用于定义事件。
- 协变量 (Covariate) 栏: 选入协变量。
- 层 (Strata) : 选入分层变量。



图 17-29 “Cox 回归”对话框



图 17-30 “事件定义 (Define Event)”对话框

2. 协变量设置

协变量 (Covariate) 栏用于选入协变量, 在其下有一 “方法 (Method)” 下拉表, 此表用于指定协变量进入回归模型的方法, 如图 17-31 所示。

- 输入 (Enter) : 所有自变量强制进入回归方程, 如果自变量较少, 则选择此项。
- 向前: 有条件 (Forward: Conditional) : 以假设参数为基础似然比概率检验, 向前逐步选择自变量。
- 向前: LR (Forward: LR) : 以最大局部似然为基础似然比概率检验, 向前逐步选择自变量。
- 向前: 瓦尔德 (Forward: Wald) : Wald 概率统计法, 向前逐步选择自变量。
- 向后: 有条件 (Backward: Conditional) : 以假设参数为基础似然比概率检验, 向后逐步选择自变量。
- 向后: LR (Backward: LR) : 以最大局部似然为基础似然比概率检验, 向后逐步选择自变量。
- 向后: 瓦尔德 (Backward: Wald) : Wald 概率统计法, 向后逐步选择自变量。

3. 分类 (Categorical) 设置

变量选入到分类 (Categorical) 栏以后, 则可以进行分类 (Categorical) 选项设置, 即分类协变量设置。单击图 17-29 右上角的 “分类 (Categorical)” 按钮, 则弹出如图 17-32 所示的对话框, 各个选项栏功能如下所述。



图 17-31 “方法 (Method)” 设置



图 17-32 “分类(Categorical)设置”对话框

协变量 (Categorical) 栏：用于存放选入的所有分类协变量。

分类协变量 (Categorical Covariates) 栏：用于选入指定为分类变量的协变量，变量名后的括号里显示的是正在使用的对照方法。

更改对比 (Change Contrast) 栏：此栏用于设置对指定协变量的对照方式，修改后，可以单击“变化量 (Change)”按钮确认。Contrast 下拉菜单有 7 种对照方式，具体如下所示。

- 指示符 (Indicator)：用于指示是否属于某一个分类。
- 简单 (Simple)：用于预测变量的每个分类都与参考分类进行比较。
- 差值 (Difference)：除了第一类外，预测变量的每个分类都与前面所有分类的平均效应进行比较。
- 赫尔默特比较：除了最后一类外，预测变量的每个分类都与其后面的所有分类的平均效应进行比较。
- 重复比较 (Repeated)：除了第一类外，预测变量的每个分类都与前面所有分类进行比较。
- 多项式 (Polynomial)：此方法假设各类别间距相等，仅适用于数值型变量。
- 偏差 (Deviation)：预测变量的每个分类都与总体效应进行比较。

参考类别 (Reference Category)：此栏用于指定参考分类。如果选择了指示符 (Indicator)、简单 (Simple)、偏差 ((Deviation) 方法，则需要指定一个参考类别，可以选择 First (第一类) 和 Last (最后一类)，系统默认为 Last。

4. 图 (Plots) 设置

图 (Plots) 选项用于图形设置，单击图 17-29 中右上角的“图 (Plots)”按钮，则系统弹出如图 17-33 所示对话框，各栏功能如下所述。

图类型 (Plot Type) 栏：用于选择输出的图形类型。

- 生存分析 (Survival)：线性刻度的累积生存函数曲线。
- 一减去生存函数 (One minus Survival)：1 减去生存函数所得的曲线。
- 风险 (Hazard)：线性刻度的累积风险函数的散点图。
- 负对数累积生存函数的对数 (Log Survival)：对数累积生存函数曲线。

协变量值的绘制位置 (Covariate Values Plotted at) 栏：此栏只有在指定了协变量为固定值时，才可以绘制生存函数关于时间的图形。默认情况时，此固定变量取值为协变量

的均值 (Mean)。如果要修改此值,则选中变量,然后在“更改值 (Change Value)”选项栏中选择“值 (Value)”选项,在其后填入相应的固定值即可,并单击“更改 (Change)”按钮确认。

针对下列各项绘制单独的线条 (Separate Lines for) 栏:用于选入一个分类协变量,绘制图形时将它作为分线变量,对其每一个取值分别绘制一条曲线。

5. 保存 (Save) 设置

单击图 17-29 中右上角的“保存 (Save)”按钮,则弹出如图 17-34 所示的对话框,此对话框用于设置保存选项,各栏选项含义如下所述。

保存模型变量 (Save Model Variables) 栏:此栏用于保存模型变量,各选项功能如下所述。

- 生存分析函数 (Survival Function):生存函数估计值。
- 风险函数 (Hazard Function):累积风险函数估计值。
- 生存分析函数的标准误差 (Standard Error of Survival Function):生存函数估计值的标准误差。
- 偏残差 (Partial Residuals)。
- 生存分析函数负对数累积函数的对数 (Log Minus Log Survival Function):对数转换后的累积生存函数。
- DiBeta (s):剔除某个观测后引起的参数估计值的变化,对最终模型的每个协变量都生成一个新变量用于保存。
- X*Beta:保存线性预测的得分,由中心化协变量与估计参数相乘后再求和所得。

将模型信息导出到 XML 文件 (Export Model Information to XML File) 栏:用于把模型信息导入 XML 文件之中,单击“浏览 (Browse)”按钮可以指定路径。



图 17-33 “Cox 回归:图”对话框



图 17-34 “Cox 回归:保存新变量”对话框

6. 选项 (Options) 设置

单击图 17-29 所示的“选项 (Options)”按钮,则弹出如图 17-35 所示的对话框,此界面主要设置一些统计量。

模型统计 (Model Statistics) 栏:此栏用于设置模型统计量,可以选择的选项如下。

- Exp (B) 的置信区间:默认区间是 95%。

- 估算值的相关性 (Correlation of Estimates) : 表示系数估计值的相关矩阵。
- 显示模型信息 (Display Model Information) 栏: 用于指定输出方式。“在每个步骤 (At Each Step)”表示逐步回归的每一步都输出相关的统计量;“在最后一个步骤 (At Last Step)”表示只在逐步回归分析的最后一步输出相关的统计量。

步进概率 (Probability for Stepwise) 栏: 用于指定协变量进入或剔除模型的临界概率。其中“进入 (Entry)”输入指定变量进入模型的概率, 默认为 0.05;“除去 (Removal)”表示指定变量剔除出模型的临界值, 默认为 0.10。

最大迭代次数 (Maximum Iterations) 栏: 用于指定最大的迭代次数, 默认为 20。

显示基线函数 (Display Baseline Function) 栏: 生成基准危险函数、协变量均值生存函数和危险函数表。



图 17-35 “Cox 回归: 选项”对话框

17.4.3 实例分析



结果文件

——附带光盘“PROGRAM\CH17\实例 17-3”文件夹



动画演示

——附带光盘“AVI\实例 17-3.avi”文件

本实例还是使用 SPSS 自带的数据集 telco.sav 来进行 Cox 回归模型的分析。此数据集为某电信公司的客户数据集, 其 SPSS 数据格式如图 17-5 所示。下面就利用 Cox 回归模型来分析客户流失问题。

1. 参数设置

选择菜单“分析 (Analyze) 生存分析 (Survival) Cox 回归 (Cox Regression)”命令, 则弹出如图 17-36 所示的“Cox 回归 (Cox Regression) 参数设置”对话框。选中变量 Months with service (tenure) 到“时间 (Time)”变量框中, 选择变量 Churn (1) 到“状态 (Status)”变量框中, 然后激活“状态 (Status)”选项栏下的“定义事件 (Define Event)”按钮。

单击“定义事件 (Define Event)”按钮, 则弹出如图 17-37 所示对话框。在“单值 (Single Value)”变量框中填入 1, 然后单击“继续 (Continue)”按钮返回主界面。

然后把变量 ed、employ、retire、gender、reside 选入到“协变量 (Covariates)”变量框中, 其下的“方法 (Method)”下拉菜单中选中“向前: LR (Forward: LR)”方法。接着单击“下一张 (Next)”按钮, 在变量 Customer category (custcat)到“协变量 (Covariates)”变量框

中,其下的“方法(Method)”下拉菜单中选中“输入(Enter)”方法。

选择单击“分类(Categorical)”按钮,弹出如图 17-38 所示的对话框,把变量 ed、retire、gender,以及 custcat 选入“分类协变量(Categorical Covariates)”变量框中,单击“继续(Continue)”按钮返回主界面。

单击“图(Plots)”按钮,则弹出如图 17-39 所示的对话框,选中“生存分析(Survival)”和“风险(Hazard)”选项,把变量 custcat 选入“针对下列各项绘制单独的线条(Separate Lines for)”变量框中,单击“继续(Continue)”按钮返回主界面。



图 17-36 “Cox 回归”对话框

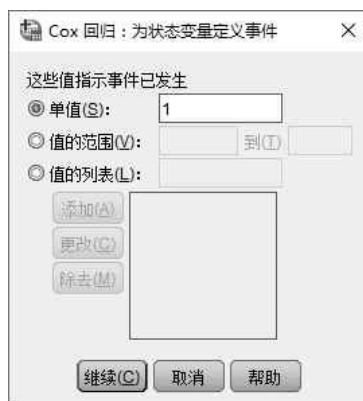


图 17-37 “Cox 回归: 为状态变量定义事件”对话框



图 17-38 “Cox 回归: 定义分类协变量”对话框



图 17-39 “Cox 回归: 图”对话框

2. 结果分析

单击 Cox 回归对话框主界面中的“确定(OK)”按钮,进行 Cox 回归分析,结果如下。首先是个案的处理简要结果,给出了数据的简要统计信息,如图 17-40 所示。

| 个案处理摘要 | | | |
|-----------------------------|-----------------------|------|--------|
| | | 个案数 | 百分比 |
| 可以在分析中使用的个案 | 事件 ^a | 274 | 27.4% |
| | 剔除后 | 726 | 72.6% |
| | 总计 | 1000 | 100.0% |
| 已删除的个案 | 具有缺失值的个案 | 0 | 0.0% |
| | 具有负时间的个案 | 0 | 0.0% |
| | 层中最早发生的事件之前 剔除后的个案 | 0 | 0.0% |
| | 总计 | 0 | 0.0% |
| 总计 | | 1000 | 100.0% |
| a. 因变量: Months with service | | | |

图 17-40 个案的处理简要结果

然后是分类变量编码，如图 17-41 所示，给出了分类变量自动编码的结果，有助于解释分类协变量的回归系数。

| 分类变量编码 ^{a,c,e,f} | | | | | | |
|---|--------------------------------|-----|------------------|-----|-----|-----|
| | | 频率 | (1) ^d | (2) | (3) | (4) |
| Level of education ^b | 1=Did not complete high school | 204 | 1 | 0 | 0 | 0 |
| | 2=High school degree | 287 | 0 | 1 | 0 | 0 |
| | 3=Some college | 209 | 0 | 0 | 1 | 0 |
| | 4=College degree | 234 | 0 | 0 | 0 | 1 |
| | 5=Post-undergraduate degree | 66 | 0 | 0 | 0 | 0 |
| Retired ^b | .00=No | 953 | 1 | | | |
| | 1.00=Yes | 47 | 0 | | | |
| Gender ^b | 0=Male | 483 | 1 | | | |
| | 1=Female | 517 | 0 | | | |
| Customer category ^b | 1=Basic service | 266 | 1 | 0 | 0 | |
| | 2=E-service | 217 | 0 | 1 | 0 | |
| | 3=Plus service | 281 | 0 | 0 | 1 | |
| | 4=Total service | 236 | 0 | 0 | 0 | |
| a. 类别变量: Level of education (ed) b. 指示符参数编码 c. 类别变量: Retired (retire) d. 由于 (0,1) 变量已重新编码，因此其系数不会与指示符 (0,1) 编码的系数相同。 e. 类别变量: Gender (gender) f. 类别变量: Customer category (custcat) | | | | | | |

图 17-41 分类变量编码

如图 17-42 所示的是变量的系数检验，是向前逐步回归的系数检验结果，如果加入一个变量后 χ^2 更改量的显著性小于 0.05，则加入此变量是合理的。

| 模型系数的 Omnibus 检验 ^c | | | | | | | | | | |
|-------------------------------|----------|---------|-----|------|----------|-----|------|----------|-----|------|
| 步长(T) | -2 对数似然 | 总体 (得分) | | | 从上一步进行更改 | | | 从上一块进行更改 | | |
| | | 卡方 | 自由度 | 显著性 | 卡方 | 自由度 | 显著性 | 卡方 | 自由度 | 显著性 |
| 1 ^a | 3357.318 | 131.930 | 1 | .000 | 169.046 | 1 | .000 | 169.046 | 1 | .000 |
| 2 ^b | 3344.089 | 147.561 | 5 | .000 | 13.229 | 4 | .010 | 182.275 | 5 | .000 |

a. 在步骤号 1: Years with current employer 处输入的变量
b. 在步骤号 2: Level of education 处输入的变量
c. 起始块号 1。方法 = 向前步进 (似然比)

图 17-42 变量的系数检验

如图 17-43 所示的是向前逐步回归估计完成以后，模型中取舍的变量情况，经过两步迭代，保留 employ 和 ed 变量。

| 方程中的变量 | | | | | | | |
|--------|-----------------------------|-------|------|---------|-----|------|--------|
| | | B | SE | 瓦尔德 | 自由度 | 显著性 | Exp(B) |
| 步骤 1 | Years with current employer | -.100 | .009 | 118.894 | 1 | .000 | .905 |
| 步骤 2 | Level of education | | | 12.820 | 4 | .012 | |
| | Level of education(1) | -.535 | .262 | 4.165 | 1 | .041 | .586 |
| | Level of education(2) | -.367 | .229 | 2.569 | 1 | .109 | .693 |
| | Level of education(3) | -.106 | .231 | .209 | 1 | .647 | .900 |
| | Level of education(4) | .080 | .216 | .135 | 1 | .713 | 1.083 |
| | Years with current employer | -.094 | .009 | 102.137 | 1 | .000 | .911 |

图 17-43 变量保留结果

图 17-44 输出的是 Enter 方法的系数检验结果，从 χ^2 更改量的显著性远小于 0.01 可以看出，加入了变量 custcat 是合理的。

| 模型系数的 Omnibus 检验 ^a | | | | | | | | | |
|-------------------------------|---------|-----|------|----------|-----|------|----------|-----|------|
| -2 对数似然 | 总体 (得分) | | | 从上一步进行更改 | | | 从上一块进行更改 | | |
| | 卡方 | 自由度 | 显著性 | 卡方 | 自由度 | 显著性 | 卡方 | 自由度 | 显著性 |
| 3312.741 | 176.815 | 8 | .000 | 31.348 | 3 | .000 | 31.348 | 3 | .000 |

a. 起始块号 2。方法 = 输入

图 17-44 Enter 方法的系数检验结果

图 17-45 输出的是 Enter 方法估计完成以后模型中取舍的变量情况，此时保留了变量 employ、ed 和 custcat。从图 17-45 中的结果可以看出，不在方程中的变量的检验结果的显著性值均大于 0.1。

如图 17-46 所示的是生存函数图形，可以看出基本服务和所有服务客户的生存函数曲线偏低。

如图 17-47 所示的是危险函数图形，此图是按照客户类型分组后的累积危险函数图形。

| 方程中的变量 | | | | | | |
|-----------------------------|-------|------|--------|-----|------|--------|
| | B | SE | 瓦尔德 | 自由度 | 显著性 | Exp(B) |
| Level of education | | | 12.935 | 4 | .012 | |
| Level of education(1) | -.613 | .278 | 4.843 | 1 | .028 | .542 |
| Level of education(2) | -.357 | .237 | 2.265 | 1 | .132 | .700 |
| Level of education(3) | -.148 | .236 | .394 | 1 | .530 | .862 |
| Level of education(4) | .102 | .217 | .221 | 1 | .638 | 1.107 |
| Years with current employer | -.090 | .009 | 93.480 | 1 | .000 | .914 |
| Customer category | | | 31.084 | 3 | .000 | |
| Customer category(1) | .325 | .168 | 3.742 | 1 | .053 | 1.384 |
| Customer category(2) | -.486 | .170 | 8.204 | 1 | .004 | .615 |
| Customer category(3) | -.530 | .195 | 7.398 | 1 | .007 | .589 |

| 未包括在方程中的变量 ^a | | | |
|-------------------------------|-------|-----|------|
| | 得分 | 自由度 | 显著性 |
| Retired | 2.227 | 1 | .136 |
| Gender | .008 | 1 | .929 |
| Number of people in household | 1.967 | 1 | .161 |

a. 残差卡方 = 4.695, 自由度为 3, 显著性 = .196

图 17-45 模型变量取舍

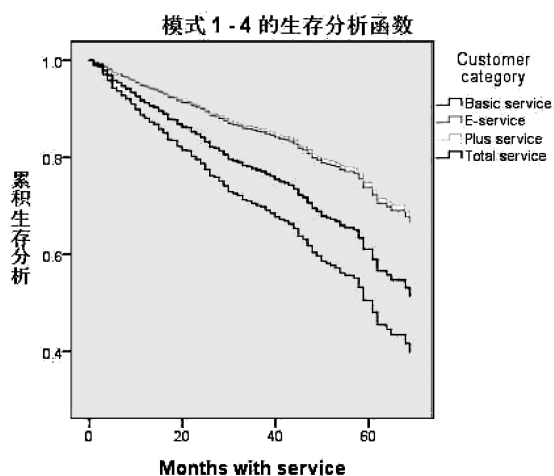


图 17-46 生存函数图形

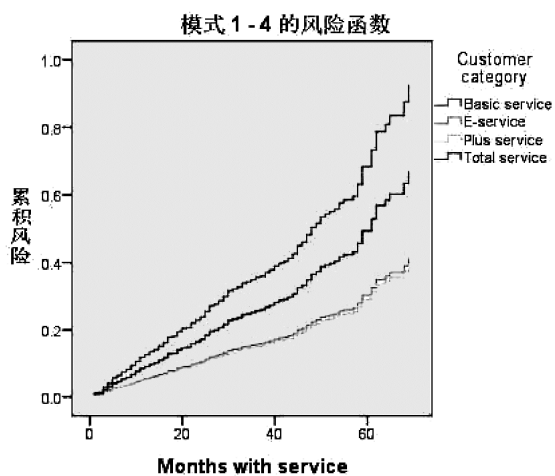
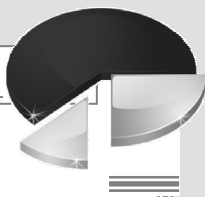


图 17-47 危险函数图形



第 18 章 对数线性模型

对数线性模型是用于离散型数据或整理成列联表格式的计数资料的统计分析工具。在对数线性模型中,所有用作分类的因素均为独立变量,列联表各单元中的例数为应变量。对于列联表资料,通常做 χ^2 检验,但 χ^2 检验无法系统地评价变量间的联系,也无法估计变量间相互作用的大小,而对数线性模型是处理这些问题的最佳方法。

本章将详细叙述 SPSS 系统中解决对数线性模型的三个过程,即 General 过程、Logit 过程,以及模型(Model) Selection 过程。



本讲内容

- 对数线性模型概述
- General 过程
- Logit 过程
- 模型(Model) Selection 过程

18.1 对数线性模型概述

对数线性模式主要目的如下。

- 提出研究模式,以计算期望次数,而且需使期望次数与实际观察次数之间没有显著差异。
- 希望以最精简的模式达到第一点的目标。
- 因变量一定要与其他变量有交互作用。

现在简单直观地通过二维表介绍一下对数线性模型,假设不同的行代表第一个变量的不同水平,而不同的列代表第二个变量的不同水平。用 m_{ij} 代表二维列联表第 i 行,第 j 列的频数。人们常假设这个频数可以用下面的公式来确定。

$$\ln m_{ij} = \alpha_i + \beta_j + \varepsilon_{ij}$$

这就是多项分布对数线性模型。这里 α_i 为行变量的第 i 个水平对 $\ln m_{ij}$ 的影响,而 β_j 为列变量的第 j 个水平对 $\ln m_{ij}$ 的影响,这两个影响称为主效应; ε_{ij} 代表随机误差。

这个模型看上去和回归模型很像，但由于对于分布的假设不同，不能简单地用线性回归的方法来套用（和 Logistic 回归类似）；计算过程也很不一样，当然我们把这个留给计算机去操心了。只要利用数据来拟合这个模型就可以得到对于参数 μ 的估计（没有意义），以及对 α_i 和 β_j 的估计。

有了“估计”的参数，就可以预测出任何 i, j 水平组合的频数 m_{ij} 了（通过其对数）。

注意，这里的估计之所以打引号是因为一个变量的各个水平的影响是相对的，因此，只有事先固定一个参数值（ $\alpha = 0$ ），或者设定类似于 $\sum \alpha_i = 0$ 这样的约束，才可能估计出各值。没有约束，则这些参数是估计不出来的。

二维列联表的更完全的对数线性模型为

$$\ln m_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$$

这里的 $(\alpha\beta)_{ij}$ 代表第一个变量的第 i 个水平和第二个变量的第 j 个水平对 $\ln m_{ij}$ 的共同影响（交叉效应）。即当单独作用时，每个变量的一个水平对 $\ln m_{ij}$ 的影响只有 α_i （或 β_j ）大，但如果这两个变量一同影响就不仅是 $+\beta_j$ ，而且还多出一项。

这里交叉项的诸参数的大小是相对的，也需要约束条件来得到其“估计”；涉及的变量和水平越多，约束也越多。

注意：无论对模型假设了多少种效应，并不见得都有意义；有些可能是多余的。本来没有交叉影响，但如果写入，也没有关系，在分析过程中一般可以知道哪些影响是显著的，而哪些是不显著的。

18.2 常规模型（General）过程

调用该过程可对一个或多个二维列联表资料进行非层次对数线性分析。它只能拟合全饱和模型，即分类变量各自效应及其相互间效应均包含在对数线性模型中。

18.2.1 常规模型分析过程的参数设置

选择菜单“分析（Analyze）对数线性模型（Loglinear）常规模型（General）”，则弹出如图 18-1 所示的对话框，对话框各个组成部分如下所述。

1. 变量选择设置

图 18-1 中的左边是待分析的变量列表框，其他选项栏功能如下。

- 因子（Factor(s)）选项栏：用于选入需要分析的因素变量，最多可以选择 10 个。
- 单元格协变量（Cell Covariate(s)）选项栏：用于选入单元格协变量。
- 单元格结构（Cell Structure）选项栏：用于选入单元格结构变量。
- 对比变量（Contrast Variable(s)）选项栏：用于选入对照变量，用于计算广义对数比率。

2. 单元格计数分布（Distribution of Cell Counts）选项栏

用于指定单元格频数的分布。

- 泊松分布。
- 多项式分布 (Multinomial)。

3. 保存 (Save) 设置

单击图 18-1 中的“保存 (Save)”按钮, 则弹出如图 18-2 所示的对话框, 此对话框用于设置保存参数, 各个组成部分如下。

- 残差 (Residuals)。
- 标准化残差 (Standardized Residuals)。
- 调整后残差 (Adjusted Residuals)。
- 偏差残差 (Deviance Residuals)。
- 预测值 (Predicted Values)。



图 18-1 “常规 (General) 设置”对话框



图 18-2 “保存 (Save) 设置”对话框

4. 模型 (Model) 设置

单击图 18-1 中的“模型 (M)”按钮, 则弹出如图 18-3 所示的对话框, 用于设置模型的参数, 各个组成部分含义如下。

指定模型 (Specify Model) 选项栏: 用于指定模型的类别。

- 饱和 (S): 指定模型包括所有因素变量的主效应和交互效应, 不包括与协变量有关的效应。
- 定制 (C): 用户自定义。

因子与协变量 (Factors & Covariates) 选项框: 用于显示因素变量的变量名称。

构建项 (Build Term(s)) 选项栏: 用于设置效应类型, 包括如下几种:

- 交互 (Interaction): 交互效应。
- 主效应 (Main Effects)。
- N 维交互效应 (All n-Way): 所有 n 维交互效应。



图 18-3 “模型 (Model) 设置”对话框

5. 选项 (Options) 设置

单击图 18-1 中的“选项 (O)”按钮，则弹出如图 18-4 所示的对话框，用于设置模型的参数，各个组成部分含义如下。

显示选项栏：用于设置哪些输出统计量。

- 频率 (Frequencies)
- 残差 (Residuals)
- 设计矩阵 (Design Matrix)
- 估算值 (Estimates)
- 迭代历史记录 (Iteration History)

图选项栏：用于设置输出图形的有关参数。

- 调整后残差 (Adjusted Residuals)
- 调整后残差的正态概率 (Normal Probability)
- 偏差残差图 (Deviance Residuals)
- 偏差残差的正态概率 (Normal probability for deviance)

置信区间选项栏：指定参数估计的置信区间。

条件选项栏：用于设置与迭代有关的参数。

- 最大迭代次数 (Maximum Iterations)
- 收敛性 (Convergence)
- Delta：指定调整系数。



图 18-4 “选项设置”对话框

18.2.2 实例分析



结果文件

——附带光盘“PROGRAM\CH18\实例 18-1”文件夹



动画演示

——附带光盘“AVI\实例 18-1.avi”文件

本实例所用数据集为 SPSS 自带的文件 demo.sav。此数据集是某公司的用户调查数据集，包括 29 个变量，6400 个观测个数，主要是关于数据库营销的数据集，数据集的格式如图 18-5 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|---------|----|----|------|---------------------|------------------|----|---|----|----|----|
| 1 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 2 | marital | 数字 | 4 | 0 | Marital status | {0, Unmarrie... | 无 | 8 | 右 | 标度 | 输入 |
| 3 | address | 数字 | 4 | 0 | Years at curren... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 4 | income | 数字 | 8 | 2 | Household inco... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 5 | inccat | 数字 | 8 | 2 | Income categor... | {1.00, Under... | 无 | 8 | 右 | 有序 | 输入 |
| 6 | car | 数字 | 8 | 2 | Price of primary... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 7 | carcat | 数字 | 8 | 2 | Primary vehicle... | {1.00, Econ... | 无 | 8 | 右 | 有序 | 输入 |
| 8 | ed | 数字 | 4 | 0 | Level of education | {1, Did not c... | 无 | 8 | 右 | 标度 | 输入 |

图 18-5 数据集格式

1. 参数设置

选择菜单“分析 (Analyze) 对数线性模型 (Loglinear) 常规模型 (General)”，则弹出如图 18-6 所示的对话框，选择变量 Newspaper subscription 和 Response 到“因子 (Factor(s))”选项栏中。

然后单击图 18-6 中的“模型 (M)”按钮，弹出如图 18-7 所示对话框，选择“定制 (C)”选项栏，选择“构建项 (Build Term(s))”下拉菜单中的“主效应 (Main Effects)”选项，然后选择变量 news 和 response 到“模型中的项 (Terms in Model)”选项栏中，然后单击“继续”按钮返回主界面。



图 18-6 “常规 (General) 设置”对话框



图 18-7 “模型 (Model) 设置”对话框

2. 结果分析

设置好上述的参数以后，单击主界面中的“确定 (OK)”按钮运行，分析结果如图 18-8 所示，它反映了分析的数据信息和模型收敛的情况。从图中可以看出共有 6400 个观测数量，无缺失值。模型的收敛信息为最大迭代次数 20，收敛公差 0.001，迭代次数 5 次。

然后是拟合优度的检验结果。如图 18-9 所示，从图中可以看出似然比和 Pearson 卡方统计量的显著性值均远远小于 0.05，所以拒绝原假设，故报纸的订阅变量（Newspaper Subscription）和是否反馈（Response）有显著的相关关系。

| 数据信息 | | | 收敛信息 ^{a,b} | |
|------|------------------------|------|--|---------------------|
| | | 个案数 | 最大迭代次数 | 20 |
| 个案 | 有效 | 6400 | 收敛容差 | .00100 |
| | 缺失 | 0 | 最终最大绝对差值 | .00032 ^c |
| | 加权有效 | 6400 | 最终最大相对差值 | .00015 |
| 单元格 | 定义的单元格 | 4 | 迭代次数 | 5 |
| | 结构零 | 0 | a. 模型：泊松 b. 设计：常量 + news + response c. 由于参数估算值的最大绝对变化量小于指定的收敛条件，因此迭代已收敛。 | |
| | 抽样零 | 0 | | |
| 类别 | Newspaper subscription | 2 | | |
| | Response | 2 | | |

图 18-8 数据信息和模型收敛信息

| 拟合优度检验 ^{a,b} | | | |
|----------------------------|--------|-----|------|
| | 值 | 自由度 | 显著性 |
| 似然比 | 49.502 | 1 | .000 |
| 皮尔逊卡方 | 50.031 | 1 | .000 |
| a. 模型：泊松 | | | |
| b. 设计：常量 + news + response | | | |

图 18-9 模型拟合优度检验结果

图 18-10 是单元格计数和残差输出结果。给出了残差、标准化残差、调整残差，以及偏差等信息，此表用来计算拟合优度统计量。

| 单元格计数和残差 ^{a,b} | | | | | | | | | |
|----------------------------|----------|------|-------|----------|-------|---------|--------|--------|--------|
| Newspaper subscription | Response | 实测 | | 期望 | | 残差 | 标准化残差 | 调整后残差 | 偏差 |
| | | 计数 | % | 计数 | % | | | | |
| Yes | Yes | 380 | 5.9% | 293.668 | 4.6% | 86.332 | 5.038 | 7.072 | 4.817 |
| | No | 2388 | 37.3% | 2474.333 | 38.7% | -86.333 | -1.736 | -7.072 | -1.746 |
| No | Yes | 299 | 4.7% | 385.333 | 6.0% | -86.333 | -4.398 | -7.072 | -4.580 |
| | No | 3333 | 52.1% | 3246.668 | 50.7% | 86.333 | 1.515 | 7.072 | 1.509 |
| a. 模型：泊松 | | | | | | | | | |
| b. 设计：常量 + news + response | | | | | | | | | |

图 18-10 单元格计数和残差的输出结果

最后输出的是调整残差的标准 Q-Q 图，以及调整残差的消除趋势标准 Q-Q 图，如图 18-11、图 18-12 所示。从图中可以看出残差存在一定的趋势，所以不服从正态分布，所用的拟合模型不能完全解释单元格频数的分布信息。

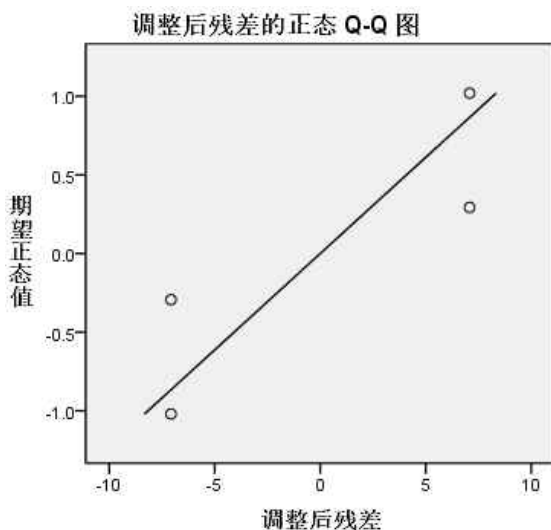


图 18-11 调整残差的标准 Q-Q 图

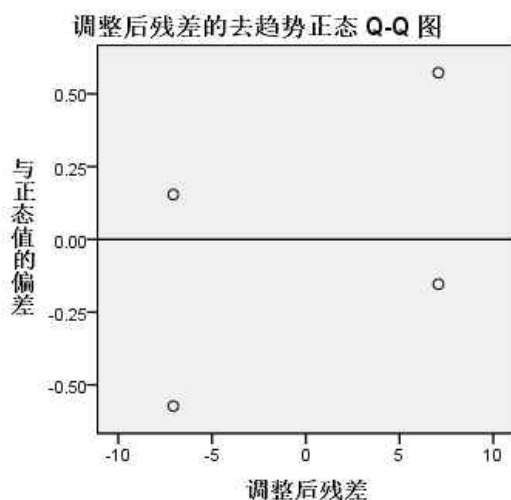


图 18-12 调整残差的消除趋势标准 Q-Q 图

18.3 分对数 (Logit) 过程

调用此过程可完成对一个应变量与一个或多个自变量之间对数线性模型的拟合。如果分类变量未区分应变量和自变量,那么应采用本章 18.1 节、18.2 节介绍的方法;如果应变量是二分计量,自变量是连续计量,那么应采用 Logistic 回归方法。

18.3.1 分对数分析过程的参数设置

选择菜单“分析 (Analyze) 对数线性模型 (Loglinear) 分对数 (Logit)”,则弹出如图 18-13 所示的对话框,对话框各个组成部分含义如下所述。

1. 变量选择设置

图 18-13 中的左边是待分析的变量列表框，其他选项栏功能如下所述。

- 因变量 (Dependent)：用于选入因变量，必须为分类变量。
- 因子 (Factor(s)) 选项栏：用于选入需要分析的因素变量。
- 单元格协变量 (Cell Covariate(s)) 选项栏：用于选入单元格协变量。
- 单元格结构 (Cell Structure) 选项栏：用于选入单元格结构变量。
- 对比变量 (Contrast Variable(s)) 选项栏：选入对照变量，以计算广义对数比率。

2. 保存 (Save) 设置

单击图 18-13 中的“保存 (Save)”按钮，则弹出如图 18-14 所示的对话框，此对话框用于设置保存参数，有如下选项。

- 残差 (Residuals)
- 标准化残差 (Standardized Residuals)
- 调整后残差 (Adjusted Residuals)
- 偏差残差 (Deviance Residuals)
- 预测值 (Predicted Values)



图 18-13 “分对数(Logit)设置”对话框



图 18-14 “保存 (Save) 设置”对话框

3. 模型 (Model) 设置

单击图 18-13 中的“模型 (Model)”按钮，则弹出如图 18-15 所示的对话框，用于设置模型的参数，与上述的常规 (General) 过程的对话框一致，各个组成部分含义如下所述。

指定模型 (Specify Model) 选项栏：用于指定模型的类别。

- 饱和 (Saturated)：指定模型包括所有因素变量的主效应和交互效应，不包括与协变量有关的效应。
- 定制 (Custom)：用户自定义。

因子与协变量 (Factors & Covariates) 选项框：用于显示因素变量的变量名称。

构建项 (Build Term (s)) 选项栏：用于设置效应类型，包括如下几种。

- 交互 (Interaction)：交互效应。
- 主效应 (Main effects)。
- N 维交互效应 (All n-Way)：所有 n 维交互效应。



图 18-15 “模型 (Model) 设置”对话框

4. 选项 (Options) 设置

单击图 18-13 中的“选项 (Options)”按钮，则弹出如图 18-16 所示的对话框，用于设置模型的参数，各个组成部分含义如下。

显示 (Display) 选项栏：用于设置输出哪些统计量。

- 频数 (Frequencies)。
- 残差 (Residuals)。
- 设计矩阵 (Design Matrix)。
- 估算值 (Estimates)。
- 迭代历史记录 (Iteration History)。

图 (Plot) 选项栏：用于设置输出图形的有关参数。

- 调整后残差 (Adjusted Residuals)。
- 调整残差的正态概率 (Normal Probability)。
- 偏差残差图 (Deviance Residuals)。
- 偏差残差的正态概率 (Normal Probability for Deviance)。

置信区间 (Confidence Interval) 选项栏：用于指定参数估计的置信区间。

标准 (Criteria) 选项栏：用于设置与迭代有关的参数。

- 最大迭代次数 (Maximum Iterations)。
- 收敛性 (Convergence)。
- Delta：指定调整系数。



图 18-16 “选项设置”对话框

18.3.2 实例分析



结果文件——附带光盘“PROGRAM\CH18\实例 18-2”文件夹



动画演示——附带光盘“AVI\实例 18-2.avi”文件

本实例所用的数据集为 SPSS 系统自带的文件 cereal.sav，此数据集为某公司关于消费者对三种早餐的购买情况的调查数据集，数据集的格式如图 18-17 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|---------|----|----|------|---------------------|------------------|----|---|----|----|----|
| 1 | agecat | 数字 | 4 | 0 | Age category | {1, Under 31... | 无 | 8 | 右 | 有序 | 输入 |
| 2 | gender | 数字 | 4 | 0 | Gender | {0, Male}... | 无 | 8 | 右 | 名义 | 输入 |
| 3 | marital | 数字 | 4 | 0 | Marital status | {0, Unmarrie... | 无 | 8 | 右 | 名义 | 输入 |
| 4 | active | 数字 | 4 | 0 | Lifestyle | {0, Inactive}... | 无 | 8 | 右 | 名义 | 输入 |
| 5 | bfast | 数字 | 4 | 0 | Preferred breakf... | {1, Breakfas... | 无 | 8 | 右 | 名义 | 输入 |

图 18-17 数据文件 cereal.sav 的格式

1. 参数设置

选择菜单“分析 (Analyze) 对数线性模型 (Loglinear) 分对数 (Logit)”，则弹出如图 18-18 所示的对话框，选中变量 Preferred breakfast 到“因变量 (Dependent)”选项栏中。选中变量 Lifestyle、Gender、Age category 到“因子 (Factor(s))”选项栏中。



图 18-18 Logit 对话框中变量选择

然后单击图 18-18 中的“模型 (Model)”按钮，弹出如图 18-19 所示的对话框，选中“定制 (Custom)”选项栏，然后选择构建项 (Build Term(s)) 下拉菜单中的“主效应 (Main Effects)”选项，然后选择变量 active 和 agecat 到“模型中的项 (Model)”选项栏中，单击“继续 (Continue)”按钮返回主界面。

接着单击图 18-18 中的“选项 (Options)”按钮，弹出如图 18-20 所示的对话框，去掉所有的“图 (Plot)”选项。

2. 结果分析

设置完成以后单击图 18-18 中的“确定”按钮进行分析，结果如下，首先是如图 18-21 所示的数据信息，以及模型的收敛情况。从图中可以看出共有 880 个观测数量，无缺失值。模型的收敛信息为最大迭代次数 20，收敛公差 0.001，迭代次数 6 等。

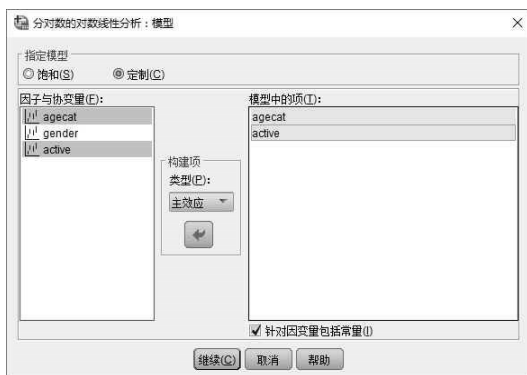


图 18-19 “模型 (Model) 设置”对话框



图 18-20 “选项 (Options) 设置”对话框

| 数据信息 | | |
|------|---------------------|-----|
| 个案数 | | |
| 个案 | 有效 | 880 |
| | 缺失 | 0 |
| | 加权有效 | 880 |
| 单元格 | 定义的单元格 | 48 |
| | 结构零 | 0 |
| | 抽样零 | 2 |
| 类别 | Preferred breakfast | 3 |
| | Age category | 4 |
| | Gender | 2 |
| | Lifestyle | 2 |

| 收敛信息 ^{a,b} | |
|---|---------------------|
| 最大迭代次数 | 20 |
| 收敛容差 | .00100 |
| 最终最大绝对差值 | .00028 ^c |
| 最终最大相对差值 | 6.50750E-5 |
| 迭代次数 | 6 |
| a. 模型: 多项分对数 | |
| b. 设计: 常量 + bfast + bfast * agecat + bfast * active | |
| c. 由于参数估计值的最大绝对变化量小于指定的收敛条件, 因此迭代已收敛。 | |

图 18-21 数据信息和模型的收敛情况

接着输出的是拟合优度检验结果, 如图 18-22 所示, 从图中可以看出似然比和 Pearson 卡方统计量的显著性值均大于 0.05, 所以接受原假设, 说明三个因素对早餐的偏好有显著影响。

| 拟合优度检验 ^{a,b} | | | |
|---|--------|-----|------|
| | 值 | 自由度 | 显著性 |
| 似然比 | 21.801 | 22 | .472 |
| 皮尔逊卡方 | 19.075 | 22 | .641 |
| a. 模型: 多项分对数 | | | |
| b. 设计: 常量 + bfast + bfast * agecat + bfast * active | | | |

图 18-22 模型的拟合优度检验结果

图 18-23 中输出的是相关性度量, 包括熵 (Entropy) 和集中度 (Concentration), 这两个统计量的取值越大, 则说明模型可以解释的离差越大, 其取值的最大值为 1。

| 相关性度量 ^{a,b} | |
|---|------|
| 熵 | .197 |
| 集中 | .189 |
| a. 模型: 多项分对数 | |
| b. 设计: 常量 + bfast + bfast * agecat + bfast * active | |

图 18-23 相关性度量

最后输出的是单元计数和残差表格,如表 18-1 所示,由于数据很多,只给出部分数据加以说明。此表给出了因素变量交叉分类的部分统计结果输出,用于计算模型的拟合优度统计量。给出的残差是观测值和期望值之间的差异,残差越小,说明模型拟合效果越好。

表 18-1 单元计数和残差表格输出

| Age category | Gender | Lifestyle | Preferred breakfast | 实 测 | | 期 望 | | 残差 | 标准化残差 | 调整残差 | 偏差 |
|--------------|--------|-----------|---------------------|-----|-------|--------|-------|--------|--------|--------|--------|
| | | | | 计数 | % | 计数 | % | | | | |
| Under 31 | Male | Inactive | Breakfast Bar | 12 | 37.5% | 11.063 | 34.6% | .937 | .348 | .405 | 1.397 |
| | | | Oatmeal | 0 | .0% | .941 | 2.9% | -.941 | -.985 | -1.130 | .000 |
| | | | Cereal | 20 | 62.5% | 19.996 | 62.5% | .004 | .001 | .002 | .089 |
| | | Active | Breakfast Bar | 28 | 52.8% | 28.553 | 53.9% | -.553 | -.152 | -.190 | -1.047 |
| | | | Oatmeal | 0 | .0% | .927 | 1.7% | -.927 | -.971 | -1.115 | .000 |
| | | | Cereal | 25 | 47.2% | 23.520 | 44.4% | 1.480 | .409 | .509 | 1.747 |
| | Female | Inactive | Breakfast Bar | 14 | 36.8% | 13.137 | 34.6% | .863 | .294 | .354 | 1.335 |
| | | | Oatmeal | 2 | 5.3% | 1.118 | 2.9% | .882 | .847 | 1.003 | 1.526 |
| | | | Cereal | 22 | 57.9% | 23.745 | 62.5% | -1.745 | -.585 | -.704 | -1.833 |
| | | Active | Breakfast Bar | 30 | 51.7% | 31.247 | 53.9% | -1.247 | -.328 | -.422 | -1.563 |
| | | | Oatmeal | 2 | 3.4% | 1.014 | 1.7% | .986 | .987 | 1.151 | 1.648 |
| | | | Cereal | 26 | 44.8% | 25.739 | 44.4% | .261 | .069 | .088 | .725 |
| 31-45 | Male | Inactive | Breakfast Bar | 16 | 37.2% | 13.955 | 32.5% | 2.045 | .666 | .793 | 2.092 |
| | | | Oatmeal | 6 | 14.0% | 6.416 | 14.9% | -.416 | -.178 | -.213 | -.897 |
| | | | Cereal | 21 | 48.8% | 22.629 | 52.6% | -1.629 | -.497 | -.594 | -1.771 |
| | | Active | Breakfast Bar | 23 | 42.6% | 28.208 | 52.2% | -5.208 | -1.419 | -1.748 | -3.064 |
| | | | Oatmeal | 8 | 14.8% | 4.948 | 9.2% | 3.052 | 1.440 | 1.670 | 2.773 |
| | | | Cereal | 23 | 42.6% | 20.845 | 38.6% | 2.155 | .602 | .731 | 2.127 |
| | Female | Inactive | Breakfast Bar | 14 | 30.4% | 14.929 | 32.5% | -.929 | -.292 | -.353 | -1.341 |
| | | | Oatmeal | 6 | 13.0% | 6.864 | 14.9% | -.864 | -.357 | -.433 | -1.270 |
| | | | Cereal | 26 | 56.5% | 24.208 | 52.6% | 1.792 | .529 | .642 | 1.927 |
| | | Active | Breakfast Bar | 37 | 58.7% | 32.909 | 52.2% | 4.091 | 1.032 | 1.330 | 2.945 |
| | | | Oatmeal | 4 | 6.3% | 5.772 | 9.2% | -1.772 | -.774 | -.925 | -1.713 |
| | | | Cereal | 22 | 34.9% | 24.319 | 38.6% | -2.319 | -.600 | -.758 | -2.100 |
| 46-60 | Male | Inactive | Breakfast Bar | 8 | 12.3% | 7.724 | 11.9% | .276 | .106 | .124 | .749 |
| | | | Oatmeal | 28 | 43.1% | 29.857 | 45.9% | -1.857 | -.462 | -.580 | -1.896 |
| | | | Cereal | 29 | 44.6% | 27.418 | 42.2% | 1.582 | .397 | .491 | 1.803 |
| | | Active | Breakfast Bar | 13 | 27.1% | 11.730 | 24.4% | 1.270 | .427 | .525 | 1.635 |
| | | | Oatmeal | 16 | 33.3% | 17.297 | 36.0% | -1.297 | -.390 | -.478 | -1.579 |
| | | | Cereal | 19 | 39.6% | 18.974 | 39.5% | .026 | .008 | .009 | .228 |
| | Female | Inactive | Breakfast Bar | 8 | 10.8% | 8.794 | 11.9% | -.794 | -.285 | -.345 | -1.230 |
| | | | Oatmeal | 38 | 51.4% | 33.991 | 45.9% | 4.009 | .935 | 1.223 | 2.911 |
| | | | Cereal | 28 | 37.8% | 31.215 | 42.2% | -3.215 | -.757 | -.971 | -2.467 |
| | | Active | Breakfast Bar | 10 | 22.7% | 10.752 | 24.4% | -.752 | -.264 | -.318 | -1.204 |
| | | | Oatmeal | 15 | 34.1% | 15.855 | 36.0% | -.855 | -.269 | -.323 | -1.290 |
| | | | Cereal | 19 | 43.2% | 17.393 | 39.5% | 1.607 | .496 | .586 | 1.833 |

续表

| Age category | Gender | Lifestyle | Preferred breakfast | 实 测 | | 期 望 | | 残差 | 标准化残差 | 调整残差 | 偏差 |
|--------------|--------|-----------|---------------------|-----|-------|--------|-------|--------|--------|--------|--------|
| | | | | 计数 | % | 计数 | % | | | | |
| Over 60 | Male | Inactive | Breakfast Bar | 1 | 1.0% | 4.581 | 4.8% | -3.581 | -1.714 | -2.059 | -1.745 |
| | | | Oatmeal | 71 | 74.0% | 70.262 | 73.2% | .738 | .170 | .225 | 1.218 |
| | | | Cereal | 24 | 25.0% | 21.158 | 22.0% | 2.842 | .700 | .918 | 2.460 |
| | | Active | Breakfast Bar | 3 | 9.1% | 3.684 | 11.2% | -.684 | -.378 | -.430 | -1.110 |
| | | | Oatmeal | 26 | 78.8% | 21.560 | 65.3% | 4.440 | 1.624 | 1.859 | 3.120 |
| | | | Cereal | 4 | 12.1% | 7.755 | 23.5% | -3.755 | -1.542 | -1.734 | -2.301 |
| | Female | Inactive | Breakfast Bar | 5 | 6.3% | 3.817 | 4.8% | 1.183 | .620 | .719 | 1.643 |
| | | | Oatmeal | 57 | 71.3% | 58.551 | 73.2% | -1.551 | -.392 | -.488 | -1.750 |
| | | | Cereal | 18 | 22.5% | 17.631 | 22.0% | .369 | .099 | .123 | .863 |
| | | Active | Breakfast Bar | 9 | 17.0% | 5.917 | 11.2% | 3.083 | 1.344 | 1.685 | 2.747 |
| | | | Oatmeal | 31 | 58.5% | 34.627 | 65.3% | -3.627 | -1.047 | -1.330 | -2.619 |
| | | | Cereal | 13 | 24.5% | 12.456 | 23.5% | .544 | .176 | .216 | 1.055 |

a. 模型：多项 Logit ；
b. 设计：常量+ bfast + bfast * agecat + bfast * active

18.4 选择模型 (Model Selection) 过程

调用该过程可对多维列联表资料进行分层对数线性分析。分层即可根据用户指定的条件，对某一或某些主效应与交互作用进行剔除，从而形成包含特定层次阶项的各种模型。

18.4.1 选择模型分析过程的参数设置

选择菜单“分析 (Analyze) 对数线性模型 (Loglinear) 选择模型 (Model Selection)”，则弹出如图 18-24 所示的对话框，对话框各个组成部分含义如下所述。

1. 变量选择设置

图 18-24 中的左边是待分析的变量列表框，其他选项栏功能如下所述。

- 因子 (Factor (s)) 选项栏：用于选入需要分析的因素变量。选入后单击其下的“定义范围 (Define Range)”按钮，则弹出如图 18-25 所示的对话框，用于指定因素变量的最小值和最大值。
- 单元格权重 (Cell Weights) 选项栏：用于选入加权变量。

2. 模型构建 (Model Building) 选项栏

用于设置模型的拟合方法。

- 使用向后去除法 (Use Backward Elimination)，在其后的最大步骤数 (Maximum Steps) 中填入最大步骤数，默认为 10，在除去概率 (Probability for Removal) 中填入指定剔除变量的临界概率，默认为 0.05。
- 一步输入 (Enter in Single Step)。



图 18-24 “选择模型对数线性分析”对话框



图 18-25 “定义范围 (Define Range) 设置”对话框

3. 模型 (Model) 设置

单击图 18-24 中的“模型 (Model)”按钮，则弹出如图 18-26 所示的对话框，用于设置模型的参数，与上述的常规 (General) 过程的对话框一致，各个组成部分含义如下。

指定模型 (Specify Model) 选项栏：用于指定模型的类别。

- 饱和 (Saturated)：指定模型包括所有因素变量的主效应和交互效应，不包括与协变量有关的效应。
- 定制 (Custom)：用户自定义。

因子与协变量 (Factors & Covariates) 选项框：用于显示因素变量的变量名称。

构建项 (Build Term(s)) 选项栏：用于设置效应类型，包括如下几种。

- 交互 (Interaction)：交互效应。
- 主效应 (Main Effects)。
- n 维交互效应 (All n -Way)：所有 n 维交互效应。

4. 选项 (Options) 设置

单击图 18-24 中的“选项 (Options)”按钮，则弹出如图 18-27 所示的对话框，用于设置模型的参数，各个组成部分含义如下。

显示 (Display) 选项栏：用于设置输出哪些统计量。

- 频率 (Frequencies)。
- 残差 (Residuals)。

图 (Plot) 选项栏：用于设置输出图形的有关参数。

- 残差：校正残差。
- 正态概率 (Normal Probability)。

饱和模型的显示 (Display for Saturated Model) 选项栏：指定饱和模型输出的统计量。

- 参数估算值 (Parameter Estimates)。
- 关联表 (Association Table)：偏相关检验表。

模型条件 (Model Criteria) 选项栏：用于设置与迭代有关的参数。

- 最大迭代次数 (Maximum Iterations)。
- 收敛性 (Convergence)。
- Delta：指定调整系数。



图 18-26 “对数线性分析：模型”对话框



图 18-27 “对数线性分析：选项”对话框

18.4.2 实例分析



结果文件——附带光盘“PROGRAM\CH18\实例 18-3”文件夹



动画演示——附带光盘“AVI\实例 18-3.avi”文件

本实例所用数据集还是 SPSS 自带的数据文件 demo.sav。此数据集是某公司的用户调查的数据集，包括 29 个变量，6400 个观测个数，主要是关于数据库营销的数据集，数据集的格式如图 18-5 所示。

1. 参数设置

选择菜单“分析 (Analyze) 对数线性模型 (Loglinear) 选择模型 (Model Selection)”命令，则弹出如图 18-28 所示的对话框，选择变量 Income category、Newspaper subscription 和 Response 到“因子 (Factor(s))”选项栏中。

然后选中 inccat，并单击其下的“定义范围 (Define Range)”按钮，则弹出如图 18-29 所示的对话框，在最小值中填入 1，在最大中填入 4。然后单击“继续 (Continue)”按钮返回主界面。

同样对话变量 news 和 response，选中后单击“定义范围 (Define Range)”按钮进行参数设置，在最小值中填入 0，在最大中填入 1。然后单击“继续 (Continue)”按钮返回主界面。

2. 结果分析

设置好上述的参数以后，单击“主界面选择模型”对话框中的“确定 (OK)”按钮进行分析，结果如下。首先是模型收敛的信息，如图 18-30 所示。



图 18-28 “选择模型对数线性分析”对话框



图 18-29 “定义范围设置”对话框

| 收敛信息 | |
|------------------|----------------------|
| 生成类 | response*news*inccat |
| 迭代次数 | 1 |
| 实测边际与拟合边际之间的最大差值 | .000 |
| 收敛条件 | 1.375 |

图 18-30 模型收敛的信息

接着输出的是模型收敛步骤的信息，如图 18-31 所示。此表给出了向后消去方法的迭代步骤，初始步骤中的三阶交互作用的卡方统计量的显著性值等于 0.262，大于 0.1，所以，从模型中删除了此三阶效应项。

| 步骤摘要 | | | | | |
|--------------------|---|-----------------|-----|------|------|
| 步骤 ^a | 效应 | 卡方 ^c | 自由度 | 显著性 | 迭代次数 |
| 0 生成类 ^b | response*news*inccat | .000 | 0 | . | |
| 删除后效应 1 | response*news*inccat | 3.998 | 3 | .262 | 4 |
| 1 生成类 ^b | response*news, response*inccat, news*inccat | 3.998 | 3 | .262 | |
| 删除后效应 1 | response*news | 67.928 | 1 | .000 | 2 |
| | 2 response*inccat | 77.562 | 3 | .000 | 2 |
| | 3 news*inccat | 224.770 | 3 | .000 | 2 |
| 2 生成类 ^b | response*news, response*inccat, news*inccat | 3.998 | 3 | .262 | |

a. 在每个步骤中，将删除“似然比变更”的显著性水平最高的效应，前提是该显著性水平大于 .050。
b. 将显示第 0 步之后的每个步骤中最佳模型的统计。
c. 对于“删除后效应”，这是将效应从模型中删除后卡方的变更。

图 18-31 模型收敛步骤的信息

最后输出的是单元计数和残差结果，如图 18-32 所示。单元计数和残差表给出了因素变量交叉分类的部分统计结果输出，用于计算模型的拟合优度统计量。给出的残差是观测值和期望值之间的差异，残差越小，说明模型拟合的效果越好。

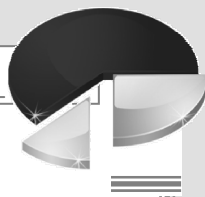
| 单元格计数和残差 | | | | | | | | |
|----------|------------------------|------------------------------|----------|-------|----------|-------|--------|-------|
| Response | Newspaper subscription | Income category in thousands | 实测 | | 期望 | | 残差 | 标准化残差 |
| | | | 计数 | % | 计数 | % | | |
| Yes | Yes | Under \$25 | 102.000 | 1.6% | 92.971 | 1.5% | 9.029 | .936 |
| | | \$25 - \$49 | 135.000 | 2.1% | 139.931 | 2.2% | -4.931 | -.417 |
| | | \$50 - \$74 | 52.000 | 0.8% | 50.229 | 0.8% | 1.771 | .250 |
| | | \$75+ | 91.000 | 1.4% | 96.853 | 1.5% | -5.853 | -.595 |
| | No | Under \$25 | 85.000 | 1.3% | 94.026 | 1.5% | -9.026 | -.931 |
| | | \$25 - \$49 | 137.000 | 2.1% | 132.068 | 2.1% | 4.932 | .429 |
| | | \$50 - \$74 | 33.000 | 0.5% | 34.772 | 0.5% | -1.772 | -.301 |
| | | \$75+ | 44.000 | 0.7% | 38.148 | 0.6% | 5.852 | .948 |
| No | Yes | Under \$25 | 319.000 | 5.0% | 328.029 | 5.1% | -9.029 | -.499 |
| | | \$25 - \$49 | 741.000 | 11.6% | 736.069 | 11.5% | 4.931 | .182 |
| | | \$50 - \$74 | 434.000 | 6.8% | 435.771 | 6.8% | -1.771 | -.085 |
| | | \$75+ | 894.000 | 14.0% | 888.147 | 13.9% | 5.853 | .196 |
| | No | Under \$25 | 668.000 | 10.4% | 658.974 | 10.3% | 9.026 | .352 |
| | | \$25 - \$49 | 1375.000 | 21.5% | 1379.932 | 21.6% | -4.932 | -.133 |
| | | \$50 - \$74 | 601.000 | 9.4% | 599.228 | 9.4% | 1.772 | .072 |
| | | \$75+ | 689.000 | 10.8% | 694.852 | 10.9% | -5.852 | -.222 |

图 18-32 单元计数和残差的输出结果

图 18-33 输出的是模型的拟合优度检验，给出了似然比和 Pearson χ^2 统计量。从表中可以看出，显著性检验的显著性值为 0.262 和 0.258，均大于 0.10，所以接受原假设，故此模型的拟合结果比较好。

| 拟合优度检验 | | | |
|--------|-------|-----|------|
| | 卡方 | 自由度 | 显著性 |
| 似然比 | 3.998 | 3 | .262 |
| 皮尔逊 | 4.029 | 3 | .258 |

图 18-33 模型的拟合优度检验结果



第 19 章 时间序列分析

时间序列是变量依相等时间间隔的顺序而形成的一系列变量值。大量社会经济统计指标都依年、季、月或日统计其指标值，随着时间的推移，形成了统计指标的时间序列。因此，时间序列是某一统计指标长期变动的数量表现。时间序列分析就是估算和研究某一时间序列在长期变动过程中所存在的统计规律性。如长期变动趋势、季节性变动规律、周期变动规律，以此预测今后的发展和变化。



本讲内容

- 时间序列概述
- 时间序列数据的预处理
- 指数平滑方法
- ARIMA 模型
- 季节性分解模型

19.1 时间序列概述

时间序列分析是一种广泛应用的数量分析方法，主要用于描述和探索现象随时间发展变化的数量规律性。时间序列分析通常分传统的时间序列分析与现代的时间序列分析两种，前者研究各种时间序列因素分解，以及长期趋势、季节变动、循环变动三要素的分析；后者则主要研究自回归（AR）模型、滑动平均（MA）模型和自回归滑动平均（ARIMA）模型。

任何事物都处于不断的运动和发展变化中，为探索现象发展变化的规律性，需要观察现象随时间变化的数量特征。把某种现象发展变化的指标数值按一定时间顺序排列起来形成的数列，称为时间序列。

19.1.1 时间序列的组成部分

事物的发展受多种因素的影响，时间序列的形成也是多种因素共同作用的结果，在一个时间序列中，有长期的起决定性作用的因素，也有临时的起非决定性作用的因素；有可

以预知和控制的因 素,也有不可预知和不可控制的因 素,这些因 素相互作用和影响,从而使时间序列变化趋势呈现不同的特点。影响时间序列的因 素大致可分为四种:长期趋势、季节变动、循环变动及不规则变动。

1. 长期趋势 (Trend)

长期趋势是指现象在相当长的一段时期内,受某种长期的、决定性的因 素影响而呈现出的持续上升或持续下降的趋势,通常以 T 表示。如中国改革开放以来国内生产总值持续上升。

2. 季节变动 (Seasonal Variation)

季节变动是指现象在一年内,由于受到自然条件或社会条件的影响而形成的以一定时期为周期 (通常指一个月或季) 的、有规则的重复变动,通常以 S 表示。如时令商品的产量与销售量,旅行社的旅游收入等都会受到季节的影响。应注意的是在这里提到的“季节”并非通常意义上的“四季”,季节变动中主要指广义的概念,可以理解为一年中的某个时间段,如一个月,一个季度,或任何一个周期。

3. 循环变动 (Cyclical Variation)

循环变动是指现象持续若干年的周期变动,通常以 C 表示。循环变动的周期长短不一,没有规律,而且通常周期较长,不像季节变动有明显的周期 (小于一年)。循环变动不是单一方向的持续变动,而是涨落相间的交替波动,如经济周期。

4. 不规则变动 (Irregular Random Variation)

不规则变动是指现象由于受偶然性因 素引起的无规律、不规则的变动,如受到自然灾害等不可抗力的影响,通常以 I 表示,这种变动一般无法作出解释。

19.1.2 时间序列的数学模型

时间序列模型分为确定性的时间序列模型和随机性的时间序列模型。

1. 确定性的时间序列模型

时间序列各影响因 素之间的关系用一定的数学关系式表示出来,就构成时间序列的分解模型,可以从时间序列的分解模型中将各因 素分离出来并进行测定,了解各因 素的具体作用如何。

通常采用加法模型和乘法模型来描述时间序列的构成。加法模型的表达式为

$$Y=T+S+C+I$$

式中, Y 表示时间序列的指标数值; T 、 S 、 C 、 I 分别表示长期趋势、季节变动、循环变动、不规则变动,使用加法模型的基本假设前提是各个影响因 素对时间序列的影响是可加的,并且是相互独立的。而乘法模型的表达式为 $Y=T \times S \times C \times I$,使用乘法模型的基本假设前提是各影响因 素对时间序列的影响是相互不独立的。

2. 随机性的时间序列模型

前面讨论了确定性时间序列模型的建立,事实上,许多现实经济现象都是通过随机时间序列模型来说明的,本节主要介绍一系列常用的时间序列模型:AR 模型、MA 模型,以及 ARIMA 模型,这类模型的建立需要较多的历史数据和较深的数学知识,实际操作必须借助计算机来完成,但是该模型在短期预测中具有较高的精度,因此,在实际中得到了广泛的应用。

平稳随机序列指如果序列 $\{y_t\}$ 二阶矩有限 ($Ey_t^2 < \infty$), 且满足如下条件。

对任意整数 t , $Ey_t = u$, u 为常数。

对任意整数 t, s , 自协方差函数 $r_{ts} = \text{cov}(y_t, y_s)$ 仅与时间间隔 $t-s$ 有关, 和起止时刻 t, s 无关, 即 $r_{ts} = r_{t-s} = r_k$ 。

则称序列 $\{y_t\}$ 为宽平稳 (或协方差平稳, 二阶矩平稳) 序列。

最简单的宽平稳过程是白噪声序列, 它是构成经济序列许多复杂过程的基石, 一般白噪声过程的定义如下。

$$E\varepsilon_t = 0。$$

$$E\varepsilon_t^2 = \sigma^2, \text{ 对所有 } t。$$

$$E\varepsilon_t\varepsilon_s = 0, t \neq s。$$

其中常见的平稳序列模型包括如下几类: AR 模型, MA 模型, ARIMA 模型。

(1) AR 模型

零均值平稳随机序列 $\{y_t\}$ 满足如下形式, 即

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (19-1)$$

式中, $\phi_1, \phi_2, \cdots, \phi_p$ 为自回归系数, 满足平稳性条件; ε_t 为白噪声序列。式 (19-1) 称为 p 阶自回归模型, 简记为 $\text{AR}(p)$ 。

(2) MA 模型

一般 MA 模型的数学形式为

$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \cdots + \phi_q \varepsilon_{t-q} \quad (19-2)$$

式中, $\phi_1, \phi_2, \cdots, \phi_q$ 为滑动平均系数; ε_t 为白噪声序列。式 (19-2) 称为 q 阶滑动平均模型, 简记为 $\text{MA}(q)$ 。

(3) ARIMA 模型

一般 ARIMA 模型的数学形式为

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \cdots + \phi_q \varepsilon_{t-q} \quad (19-3)$$

式中, $\phi_1, \phi_2, \cdots, \phi_p$ 为自回归系数, 满足平稳性条件; $\phi_1, \phi_2, \cdots, \phi_q$ 为滑动平均系数; ε_t 为白噪声序列, 式 (19-3) 称为 p 阶自回归- q 阶滑动平均模型, 简记为 $\text{ARIMA}(p, q)$ 。

从以上定义中可以看出, AR 模型和 MA 模型即为 ARIMA 模型的特例, 如下。

当 $p=0$, $\text{ARIMA}(p, q)$ —— $\text{MA}(q)$ 。

当 $q=0$, $\text{ARIMA}(p, q)$ —— $\text{AR}(p)$ 。

19.1.3 时间序列的分析步骤

一个时间序列通常存在长期趋势变动、季节变动、周期变动和不规则变动因素。时间序列分析的目的就是逐一分解和测定时间序列中各项因素的变动程度和变动规律,然后将其重新综合起来,预测统计指标今后综合的变化和发展情况。

时间序列的综合分析步骤如下。

确定时间序列的变动因素和变动类型。

计算调整月(季)指数,以测定季节变动因素的影响程度。

调整时间序列的原始指标值,以消除季节变动因素的影响。

根据调整后时间序列的指标值(简称调整值)拟合长期趋势模型。

计算趋势比率或周期余数比率,以度量周期波动幅度和周期长度。

预测统计指标今后的数值。

19.1.4 SPSS 时间序列分析功能

选择菜单“分析(Analyze) 时间序列预测(Forecast)”命令,如图 19-1 所示。

首先是“创建时间因果模型”、“创建传统模型”、“应用时间因果模型”和“应用传统模型”选项,其中“创建时间因果模型”用于发现时间序列数据中的关键因果关系,“创建传统模型”选项分为指数平滑模型和自回归滑动平均模型。“应用时间因果模型”和“应用传统模型”选项执行应用传统模型的功能。

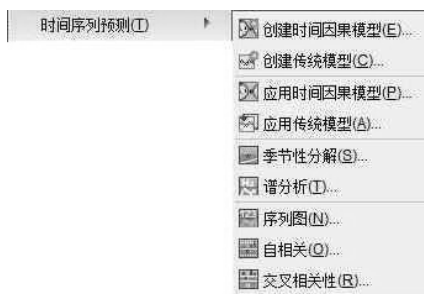


图 19-1 时间序列(Time Series)模块

时间序列预测(Time Series)提供的模型如下。

- 指数平滑法模型(Exponential Smoothing)
- 单/多变量自回归积分移动平均值模型(ARIMA)

图 19-1 中的“谱分析(Sequence Analysis)”选项实现的是谱分析功能,“序列图(Sequence Charts)”选项实现谱密度图分析功能;“自相关(Autocorrelations)”选项实现自相关图功能;“交叉相关图(Cross-Correlations)”选项实现互相关图功能。

1. 创建传统模型(Create Models)参数设置

选择菜单“Analyze 时间序列预测(Time Series) 创建传统模型(Create Models)”命令,则弹出如图 19-2 所示的对话框,此对话框用于指定分析变量、选择模型方法等操作。各选项框功能如下所述。

(1) 变量(Variables)选项栏

此栏用于设置变量,分析模型,分析数据范围等信息。

- 变量(Variables)列表:用于显示当前可用的变量。
- 因变量(Dependent Variables):用于选入因变量。
- 自变量(Independent Variables):用于选入自变量。

- 方法 (Method) : 用于指定建模方法。其中专家模型 (Expert Modeler) 表示对每个因变量分别自动寻找最优的拟合模型; 指数平滑法 (Exponential Smoothing) 用于指定指数平滑模型; ARIMA 用于指定 ARIMA 模型。
- 估算期 (Estimation Period) : 指定模型估算时的数据范围, 默认为当前所有数据。
- 预测期 (Forecast Period) : 设置预测范围, 显示要预测的数据范围。默认为当前记录的所有范围。



图 19-2 “时间序列预测建模器”对话框

(2) 统计量 (Statistics) 选项栏

此栏用于设置一些关于统计量的信息, 如图 19-3 所示。

按模型显示拟合测量、杨-博克斯统计和离群值数目 (Display fit measures, Ljung-Box statistic, and number of outlier by model) 选项。此项设置模型拟合方法、杨-博克斯统计量、由模型定义的异常点个数等, 选中后则激活其下的“拟合测量 (Fit Measures)”选项栏。

拟合测量 (Fit Measures) 选项: 此项设置输出那些反映模型拟合优度的统计量。各选项功能如下所述。

- 平稳 R 方 (Stationary R square): 平稳 R 统计量, 用于比较模型中固定成分与一个简单均值模型的差别, 当原始序列中有趋势成分或季节成分时, 优于 R 方统计量。
- R 方 (R Square): R 方统计量, 用来估计由模型解释的变异在总变异中的比例, 当原始序列为平稳序列时, 优于平稳 R 方统计量。
- 均方根误差 (Root Mean Square Error): 均方误差, 用来度量原始因变量序列与它的模型预测值的差异。

- 平均绝对误差百分比 (Mean Absolute Percentage Error): 绝对比例误差均值, 用于度量原始因变量与预测值之间的差异。
- 平均绝对误差 (Mean Absolute Error): 绝对误差均值, 用于度量原始变量与预测值之间的差异。
- 最大绝对误差百分比 (Maximum Absolute Percentage Error): 最大绝对比例误差, 以比例形式表示最大预测误差。
- 最大绝对误差 (Maximum Absolute Error): 用来度量最大的预测误差。
- 正态化 (Normalized BIC): 正态 BIC 统计量, 用来度量模型的拟合优度, 同时考虑了模型的复杂程度。



图 19-3 “统计量”选项卡

用于比较模型的统计 (Statistics for Comparing Models) 栏: 此栏用于设置模型比较的统计量输出。

- 拟合优度 (Goodness of Fit): 把每个模型的拟合优度统计量输出到一张表格里。
 - 残差自相关函数 (Residual Autocorrelation Function): 残差的自相关函数, 输出每个模型的残差自相关函数的统计特征和百分位点。
 - 残差偏自相关函数 (Residual Partial Autocorrelation Function): 残差的偏相关函数, 输出每个模型的残差偏相关函数的统计特征和百分位点。
- 单个模型的统计量: 此栏设置单个模型的输出信息。
- 参数估算值 (Parameter Estimates): 参数估计值, 对指数平滑模型和 ARIMA 模型, 分别输出它们各自的参数估计表。对于异常值的估计, 将单独输出一张表格。
 - 残差自相关函数 (Residual Autocorrelation Function): 残差的自相关函数, 输出每

个模型的残差自相关序列及其置信区间。

- 残差偏自相关函数 (Residual Partial Autocorrelation Function): 残差的偏相关函数, 输出每个模型的残差偏相关序列及其置信区间。

显示预测值 (Display Forecasts) 栏: 此项表示为每个估计模型输出预测值及其置信区间。

(3) 图 (Plots) 选项栏

单击图 19-2 中的“图 (Plots)”标签, 则弹出如图 19-4 所示的对话框, 此选项栏用于设置绘图选项设置。

用于比较模型的图 (Plots for Comparing Models): 此栏设置模型比较的图形输出。各选项与图 19-3 中拟合度量 (Fit Measure) 栏选项设置一致。

单个模型的图 (Plots for Individual Models): 此栏用于设置单个模型的绘制图形选项。其下的“序列 (Series)”选项选中后激活“每张图显示内容 (Each Plot Displays)”选项栏, 输出序列图形, 其中有 5 个选项。

- 观察值 (Observed Values): 因变量的原始观测序列。
- 预测值 (Forecasts): 预测范围的观测预测值。
- 拟合值 (Fit values): 估计范围的观测预测值。
- 预测值的置信区间 (Confidence Intervals For Forecasts): 预测范围内的置信区间。
- 拟合值的置信区间 (Confidence Intervals for Fit Values): 估计范围内的置信区间。

残差自相关函数 (Residual Autocorrelation Function): 此栏输出每个模型的残差自相关序列图。

残差偏自相关函数 (Residual Partial Autocorrelation Function): 此栏输出每个模型的残差偏相关序列图。

(4) 输出过滤 (Output Filter) 选项栏

单击图 19-2 中的“输出过滤 (Output Filter)”选项卡, 弹出如图 19-5 所示的对话框。此对话框用于设置关于输出限制选项的参数。

在输出中包括所有模型 (Include all Models in Output): 表示输出所有模型的分析结果, 这是默认选项。

根据拟合优度过滤模型 (Filter Models Based on Goodness of Fit): 此栏只输出某些模型的分析结果。

- 最佳拟合模型 (Best-fitting Models) 拟合栏表示拟合优度最好的模型, 其中模型的固定数量 (Fixed Number of Models) 表示显示的最好模型个数, 数为输出模型个数; 占模型总数的百分比 (Percentage of Total Number of Models) 表示要显示的最好模型个数占总模型个数的比例。
- 最差拟合模型 (Poorest-fitting Models) 拟合栏表示输出拟合优度最差的模型, 其中模型的固定数量 (Fixed Number of Models) 表示显示的最差模型个数, 数为输出模型个数; 占模型总数的百分比 (Percentage of Total Number of Models) 表示要显示的最差模型个数占总模型个数的比例。

拟合优度测量 (Goodness of Fit Measure) 下拉菜单, 指定衡量模型优劣的拟合优度统计量, 系统默认为平稳的 R 方 (Stantionary R square) 统计量。



图 19-4 “图表”选项卡



图 19-5 “输出过滤”选项卡

(5) 保存 (Save) 选项栏

单击“保存 (Save)”标签，弹出如图 19-6 所示的对话框，各选项框功能如下所述。

保存变量 (Save Variables)：此栏用于设置关于模型预测值的保存选项。

- 预测值 (Predicted Values) : 模型预测值。
- 置信区间下限 (Lower Confidence Limits) : 预测值的置信下限。
- 置信区间上限 (Upper Confidence Limits) : 预测值的置信上限。
- 噪声残差 (Noise Residuals) : 预测值的残差。

导出模型文件 (Export Model File) : 设置输出模型信息输出到指定的 XML 文件 , 单击 “ 浏览 (Browse) ” 按钮指定文件路径。



图 19-6 “ 保存 (Save) 设置 ” 对话框

(6) 选项 (Options) 选项栏

单击 “ 选项 (Options) ” 标签 , 弹出如图 19-7 所示对话框 , 此选项栏用于设置关于预测范围、缺失值处理方式、置信区间等选项参数。各选项功能如下所述。

预测期 (Forecast Period) 栏 : 此栏设置预测范围。

- 评估期结束后的第一个个案到活动数据集中的最后一个个案 (First case after end of Estimation Period through last case in active dataset) : 表示预测范围从估计模型所用数据的最后一个记录到当前数据集的最后一个记录。
- 评估期结束后的第一个个案到指定日期之间的个案 (First case after end of Estimation Period through a specified date) : 表示预测范围从估计模型所用数据的最后一个记录到用户指定的某个日期 , 常用来预测超过当前数据集的时间范围的记录 , 其下的日期 (Date) 栏用于指定要预测的日期。

用户缺失值 (User-Missing Values) : 设置缺失值的处理方式。

- 视为无效 (Treat as Invalid) : 表示把用户定义缺失值当做系统缺失值对待 , 作为无效数据。
- 视为有效 (Treat as Valid) : 表示把用户定义缺失值作为有效数据。

置信区间宽度 (Confidence Interval Width (%)) : 置信区间, 系统默认为 95%。

输出中的模型标识前缀 (Prefix for Model Identifiers in Output) : 指定在输出结果中用于区分不同模型的名称前缀, 默认为模型 (Model)。

ACF 和 PACF 输出中显示的最大延迟数 (Maximum Number of Lags Shown in ACF and PACF Output) : 指定自相关函数和偏相关函数的最大延迟阶数, 默认为 24。

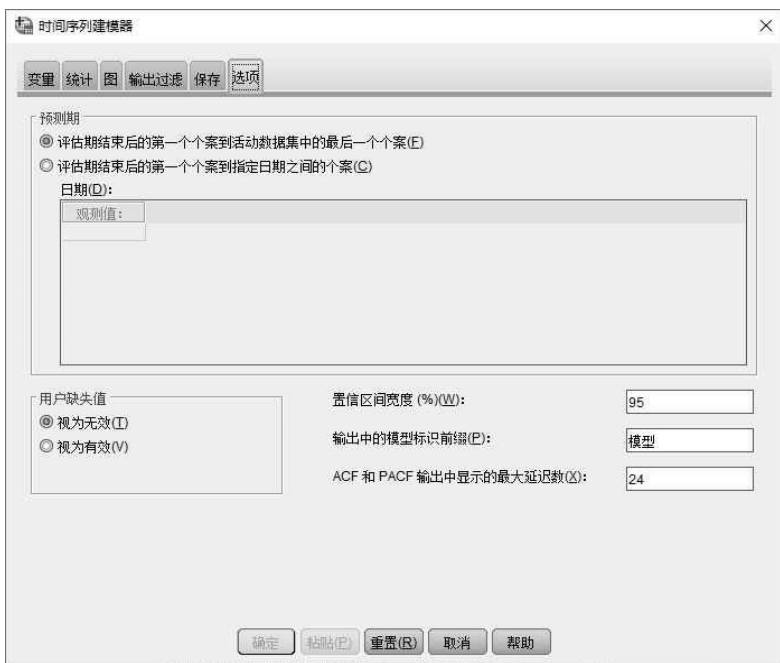


图 19-7 “选项 (Options) 设置”对话框

2. 应用传统模型 (Apply Models) 参数设置

单击选择 “Analyze 时间序列预测 (Time Series) 应用传统模型 (Apply Models)” 命令, 则弹出如图 19-8 所示的对话框, 此界面主要应用先前确定的模型, 直接对指定数据集进行分析, 各选项栏功能如下。

模型文件 (Model) 选项: 用于指定 XML 模型文件的路径和名称, 单击 “浏览 (Browse)” 按钮即可打开。

模型参数和拟合优度测量 (Model Parameters and Goodness of Fit Measures) 栏: 设置模型估计的参数和模型拟合优度统计量的引入方式。

- 从模型文件中载入 (Load from Model File): 表示从指定的模型文件中直接读取。
- 模型数据重新评估 (Reestimate from Data): 利用当前数据集重新估计模型。

预测期 (Forecast Period) 栏: 用于指定模型预测范围, 与创建传统模型 (Create Model) 选项中的设置一样。

除了模型 (Models) 选项以外, 还有统计量 (Statistics) 选项、图 (Plots) 选项、输出过滤 (Output) 选项、保存 (Save) 选项, 以及选项 (Options) 选项, 这些选项中的设置与创建传统模型 (Create Model) 选项栏中的设置基本一样, 在此不再赘述。



图 19-8 “应用时间系列模型”对话框

19.2 时间序列数据的预处理

在进行时间序列预测分析之前，一定要对原始数据集进行预处理分析，否则在随后的分析操作中可能带来不必要的麻烦和困难。本节只讲述预处理的方法，在下面几节中会进行数据预处理的实际操作设置。

19.2.1 缺失值替换

在进行时间序列预测分析之前，需要对原始数据进行初步分析，以确保数据完整。选择菜单“转换（Transform） 替换缺失值（Replace Missing Values）”命令，则系统执行缺失值替换操作，如图 19-9 所示。



图 19-9 “替换缺失值（Replace Missing Values）”对话框

各个选项框功能具体如下所述。

新变量 (New Variable (s)) : 此栏用于从变量列表中选入含有缺失值的变量。

名称和方法 (Name and Method) 选项栏 : 此栏用于设置替换缺失值的参数。其下的名称 (Name) 输入框用于指定新变量的名称。方法 (Method) 下拉菜单用于选择替换缺失值的方法, 如图 19-10 所示。

- 序列平均值 (Series Mean) : 全体序列的均值, 系统默认选项。
- 临近点的平均值 (Mean of Nearly Points) : 相邻若干点的均值。
- 临近点的中间值 (Median of nearby Points) : 相邻若干点的中位数。
- 线性插值 (Linear Interpolation) : 线性内插, 使用当前缺失值前后两个有效数据计算均值。
- 临近点处的线性趋势 (Linear Trend at Point) : 该点的线性趋势, 将记录号作为自变量, 序列值作为因变量进行回归, 求得该点的估计值。

19.2.2 定义时间变量

时间序列数据集中必须设置时间变量, 系统才能识别。选择菜单“数据 (Data) 定义日期和时间 (Define Dates)”命令, 则系统执行时间变量的功能, 如图 19-11 所示为弹出的对话框。

个案为 (Cases Are) 选项框 : 此栏给出了许多种时间格式, 用户可以进行选择。

第一个个案是 (First Case Is) 栏 : 此栏用于指定起止时间, 此栏显示的是对应个案为 (Cases Are) 选项框中不同时间格式的选项。

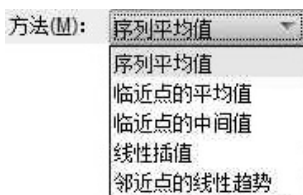


图 19-10 “方法 (Method)”选项



图 19-11 “定义日期”对话框

19.2.3 时间序列预测的平稳化

SPSS 中的创建时间序列预测 (Time Series) 过程用来对原始数据进行预处理, 包括数据差分、移动平均等操作。选择菜单“转换 (Transform) 创建时间序列预测 (Create Time Series)”命令, 则弹出如图 19-12 所示的对话框。

变量 新名称栏 : 此栏用于从变量表中选入原始的时间序列变量。

名称和函数 (Name and Function) 栏 : 用于设置对变量进行转换的参数。其下的“名称 (Name)”选项用于指定新变量的名称, 修改后单击“更改 (Change)”按钮加以确

定。“函数 (Function)” 下拉菜单用于选择转换原序列的方法, 如图 19-13 所示。

- 差异 (Difference)
- 季节性差异 (Seasonal Difference)
- 中心移动平均值 (Centered Moving Average)
- 前移动平均值 (Prior Moving Average)
- 运行中位数 (Running Medians), 以当前值计算指定范围的中位数, 下面的 Span 默认为 1。
- 累积求和 (Cumulative Sum)
- 延迟 (Lag), 其下的顺序 (Order) 框中设置滞后阶数。
- 提前 (Lead), 其下的顺序 (Order) 框中设置提前阶数。

当前周期长度 (Current) 栏: 显示当前时间变量的周期, 如果没有定义时间变量, 则此处为空。



图 19-12 “创建时间序列预测”对话框



图 19-13 模型选择设置

19.3 指数平滑模型过程

19.3.1 指数平滑的基本原理

指数平滑法是布朗 (Robert G. Brown) 提出的, 布朗认为时间序列的态势具有稳定性或规则性, 所以时间序列可被合理地顺势推延; 他认为最近的过去态势, 在某种程度上会持续到最近的未来, 所以将较大的权数放在最近的资料中。

指数平滑法是生产预测中常用的一种方法。也用于中短期经济发展趋势预测, 所有预测方法中, 指数平滑是用得最多的一种。简单的全期平均法是对时间数列的过去数据一个不漏地全部加以同等利用; 移动平均法则不考虑较远期的数据, 并在加权移动平均法中给予近期资料更大的权重; 而指数平滑法则兼容了全期平均和移动平均所长, 不舍弃过去的数据, 但是仅给予逐渐减弱的影响程度, 即随着数据的远离, 赋予逐渐收敛为零的权数。

也就是说指数平滑法是在移动平均法基础上发展起来的一种时间序列分析预测法, 它

是通过计算指数平滑值,配合一定的时间序列预测模型对现象的未来进行预测。其原理是任一期的指数平滑值都是本期实际观察值与上一期指数平滑值的加权平均。

据平滑次数不同,指数平滑法分为一次指数平滑法、二次指数平滑法和三次指数平滑法等。

1. 一次指数平滑预测

已知时间序列为 y_1, y_2, \dots, y_T , T 为序列总记录期数,一次指数平滑值为

$$S_t^{(1)} = \alpha y_t + (1 - \alpha) S_{t-1}^{(1)}, t = 1, 2, \dots, T \quad (19-4)$$

式中,上标“(1)”表示一次指数平滑; α 为平滑系数,取值为 $0 \sim 1$ 。式(19-4)表明 t 期的一次指数平滑值等于本期的实际值与上期的一次指数平滑值的加权和。指数平滑法如何克服移动平均法的不足之处,通过将 $S_t^{(1)}$ 展开即可一目了然。

$$\begin{aligned} S_t^{(1)} &= \alpha y_t + (1 - \alpha) S_{t-1}^{(1)} \\ &= \alpha y_t + \alpha(1 - \alpha) y_{t-1} + \alpha(1 - \alpha)^2 y_{t-2} + \dots + \alpha(1 - \alpha)^{t-1} y_1 + (1 - \alpha)^t S_0^{(1)} \end{aligned} \quad (19-5)$$

由式(19-5)看出, $S_t^{(1)}$ 的主要部分是 $y_t, y_{t-1}, \dots, y_2, y_1$ 的加权平均,权数由近及远分别为 $\alpha, \alpha(1 - \alpha), \alpha(1 - \alpha)^2 \dots$ 按几何级数衰减,满足近期权数大,远期权数小的要求,而且利用了时间序列的全部数据信息。由于加权系数符合指数规律,又具有平滑数据的作用,故称为指数平滑法。

(1) 选择平滑系数 α 的方法

在应用指数平滑法进行预测时,选择合适的平滑系数是非常重要的,选择是否得当直接影响到预测结果。 α 越大,说明预测越依赖于近期信息; α 越小,则表示预测更依赖于历史信息。 α 的大小,也体现了修正幅度的大小, α 越大,修正幅度越大;反之, α 越小,修正幅度也越小。一般说来, α 取值应遵循下述原则。

如果预测目标的时间序列虽然有不规则的起伏变动,但整个长期发展趋势呈比较稳定的水平趋势,则 α 应取小一些,一般可在 $0.05 \sim 0.20$ 之间取值,这时预测模型包含了较长的时间序列信息,从而使各期预测值对预测结果有相似的影响。

当时间序列波动很大,长期趋势变化幅度较大时, α 取值应大一些,可在 $0.3 \sim 0.5$ 之间选值,这时模型能迅速地根据当前的信息对预测进行大幅度修正。

当时间序列具有明显上升或下降趋势时,则 α 应取较大的值,一般取值范围为 $0.6 \sim 0.9$ 。

在实际应用中,可取若干个 α 值进行试算比较,选择预测误差最小的 α 值。

(2) 确定初始值的方法

分析以上的指数平滑公式发现,要计算指数平滑值,首先必须确定一个初始 $S_0^{(1)}$ 。由于当 $t \rightarrow 0$ 时, $S_0^{(1)}$ 的系数 $(1 - \alpha)^t \rightarrow 0$,这说明随着 t 的增大, $S_0^{(1)}$ 对预测值的影响越来越小。为计算方便,确定初始值,一般可作如下考虑:若时间序列观察期 $n > 15$ 时,以第一期观察值作为初始值;若 $n < 15$ 时,可以取最初几期的观察值的平均值作初始值,通常可取前 3 个观察期数据的平均值作为初始值。

(3) 建立预测模型的方法

如果时间序列的变化呈水平趋势,可用第 t 期的一次指数平滑值作为第 $t+1$ 期的预测值,一次指数平滑法只能用于下一期的预测,其预测模型为

$$\hat{y}_{t+1} = S_t^{(1)} = \alpha y_t + (1-\alpha)\hat{y}_t$$

上式说明, $t+1$ 期预测值是 t 期观测值和 t 期预测值的加权平均。用 y_t 代表新的数据信息,用 \hat{y}_t 代表历史的数据信息,若取 α 为 0.5,则表明预测者认为新的数据信息和历史的数据信息是同等重要;若 α 大于 0.5,表明预测者更重视新的数据信息。

也可改写为

$$\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t)$$

上式说明, $t+1$ 期预测值是在原预测值的基础上利用原预测误差进行修正得到的。 α 既代表了预测模型对时间序列的反应速度,又决定了预测模型修正误差的能力, α 的选取直接影响预测结果。

2. 二次指数平滑预测

对一次平滑的结果再进行一次平滑计算得到的数值,称为二次指数平滑值,其计算公式为

$$S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha)S_{t-1}^{(2)}, \quad t=1,2,\dots,T$$

式中, $S_t^{(2)}$ 为第 t 期的二次指数平滑值。

建立预测模型的方法

当时间序列的变动呈现出直线趋势时,用一次指数平滑方法分析仍存在着明显的滞后偏差,因此也需要修正,修正的方法是在一次指数平滑的基础上再作二次指数平滑,利用滞后偏差的规律找出曲线的发展方向和发展趋势,然后建立线性趋势预测模型。具体方法步骤如下。

第一步:确定平滑系数和初始值。方法与一次指数平滑法相同,一般取一次和二次平滑值的初始值相同;

第二步:对时间序列计算一次和二次指数平滑数值;

第三步:利用一次和二次指数平滑数值估计线性趋势模型的系数为

$$\hat{a}_t = 2S_t^{(1)} - S_t^{(2)}, \quad \hat{b}_t = \frac{\alpha}{1-\alpha}(S_t^{(1)} - S_t^{(2)})$$

这样就可以建立线性趋势预测模型 $\hat{y}_{t+m} = \hat{a}_t + \hat{b}_t m$ ($m=1,2,\dots$), 并进行预测。

3. 三次指数平滑预测

如果时间序列的变化呈现二次曲线趋势时,可用三次指数平滑法进行预测。三次指数平滑法,就是将二次指数平滑序列再进行一次指数平滑,其计算公式为

$$S_t^{(3)} = \alpha S_t^{(2)} + (1-\alpha)S_{t-1}^{(3)}$$

α 和初始值 $S_0^{(1)}, S_0^{(2)}, S_0^{(3)}$ 确定原则和方法与一次指数平滑方法相同。

三次指数平滑的目的与二次指数平滑类似,是为了计算二次曲线预测模型的参数,三次指数平滑法建立的预测模型具有多次预测能力。如果设时间序列的二次曲线预测模型为

$$\hat{y}_{t+m} = \hat{a}_t + \hat{b}_t m + \hat{c}_t m^2, \quad m=1,2,\dots$$

其中的参数计算公式分别为

$$\begin{aligned}\hat{a}_t &= 3S_t^{(1)} - 3S_t^{(2)} + S_t^{(3)} \\ \hat{b}_t &= \frac{\alpha}{2(1-\alpha)^2} [(6-5\alpha)S_t^{(1)} - 2(5-4\alpha)S_t^{(2)} + (4-3\alpha)S_t^{(3)}] \\ \hat{c}_t &= \frac{\alpha^2}{2(1-\alpha)^2} [S_t^{(1)} - 2S_t^{(2)} + S_t^{(3)}]\end{aligned}$$

19.3.2 指数平滑模型分析过程的参数设置

选择菜单“分析（Analyze）预测（Forecast）创建传统模型（Create Models）”命令，则弹出如图 19-2 所示的对话框，选择“方法（Method）”下拉菜单中指数“平滑法（Exponential Smoothing）”选项，然后再单击“条件（Criteria）”按钮，则弹出如图 19-14 所示对话框，各选项功能如下。

模型类型（Model Type）选项栏，此栏用于设置模型。

- 非季节性（Nonseasonal）：无季节因素模型，包括简单指数平滑（Simple）、霍尔特线性趋势模型（Holt's linear trend）、布朗线性趋势模型（Brown's linear trend）、衰减趋势模型（Damped trend）。
- 季节性模型（Seasonal）：包括简单季节模型（Simple seasonal）、温特斯加法模型（Winter's additive）、温特斯乘法模型（Winter's multiplicative）。

当前周期长度（Current Periodicity）栏，显示当前数据集的周期，如果没有定义，则显示为 None。

因变量转换（Dependent Variable Transformation）栏，指定因变量的变换方法。

- 无（None）：不作任何变换。
- 平方根（Square root）：平方根变换。
- 自然对数（Natural log）：自然对数变换。



图 19-14 “指数平滑条件”对话框

19.3.3 实例分析



结果文件

——附带光盘“PROGRAM\CH19\实例 19-1”文件夹



动画演示

——附带光盘“AVI\实例 19-1.avi”文件

表 19-1 是某商场电视机的销售数据，该商城希望通过一些商品近几年的销售趋势来预测未来的销售情况，以便决定下一步的促销力度和营销策略。

表 19-1 电视机的销售数据

| | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|
| 日期 | 200607 | 200608 | 200609 | 200610 | 200611 | 200612 |
| 销售额/万元 | 68.44 | 70.24 | 68.18 | 72.16 | 74.82 | 79.04 |
| 日期 | 200701 | 200702 | 200703 | 200704 | 200705 | 200706 |
| 销售额/万元 | 80.02 | 80.35 | 85.74 | 86.24 | 94.86 | 98.42 |
| 日期 | 200707 | 200708 | 200709 | 200710 | 200711 | 200712 |
| 销售额/万元 | 100.44 | 102.47 | 106.20 | 112.43 | 115.56 | 118.64 |

本案例中对电视机的销售数据进行一次指数平滑操作，以进行预测分析。

1. 对电视机的销售数据进行分析

把表 19-1 中的数据输入 SPSS 数据窗口之中，如图 19-15 所示。



图 19-15 电视机销售数据

首先对时间变量进行设置，选择菜单“数据（Data） 定义日期和时间（Define Dates）”命令，则弹出如图 19-11 所示对话框，选中变量“年份，月份（Year，months）”，在其右边的 Year 变量框和 Month 变量框中填入 2006 和 7，然后单击“确定（OK）”按钮，则进行设置完成。输出结果如图 19-16 所示。

| | 时间 | 销售额 | YEAR | MONTH | DATE |
|----|--------|--------|------|-------------|------|
| 1 | 200607 | 68.44 | 2006 | 7 JUL 2006 | |
| 2 | 200608 | 70.24 | 2006 | 8 AUG 2006 | |
| 3 | 200609 | 68.18 | 2006 | 9 SEP 2006 | |
| 4 | 200610 | 72.16 | 2006 | 10 OCT 2006 | |
| 5 | 200611 | 74.82 | 2006 | 11 NOV 2006 | |
| 6 | 200612 | 79.04 | 2006 | 12 DEC 2006 | |
| 7 | 200701 | 80.02 | 2007 | 1 JAN 2007 | |
| 8 | 200702 | 80.35 | 2007 | 2 FEB 2007 | |
| 9 | 200703 | 85.74 | 2007 | 3 MAR 2007 | |
| 10 | 200704 | 86.24 | 2007 | 4 APR 2007 | |
| 11 | 200705 | 94.86 | 2007 | 5 MAY 2007 | |
| 12 | 200706 | 98.42 | 2007 | 6 JUN 2007 | |
| 13 | 200707 | 100.44 | 2007 | 7 JUL 2007 | |
| 14 | 200708 | 102.47 | 2007 | 8 AUG 2007 | |
| 15 | 200709 | 106.20 | 2007 | 9 SEP 2007 | |
| 16 | 200710 | 112.43 | 2007 | 10 OCT 2007 | |
| 17 | 200711 | 115.56 | 2007 | 11 NOV 2007 | |
| 18 | 200712 | 118.64 | 2007 | 12 DEC 2007 | |

图 19-16 时间变量设置结果

(1) 参数设置

选择菜单“分析 (Analyze) 时间序列预测 (Forecast) 创建传统模型 (Create Models)”命令, 弹出如图 19-17 所示对话框, 把变量销售额选入“因变量 (Dependent Variables)”变量框中, 在“方法 (Method)”下拉选项栏中选中“指数平滑法 (Exponential Smoothing)”方法, 并单击其后的“条件 (Criteria)”按钮, 打开如图 19-18 所示对话框进行指数平滑的设置。

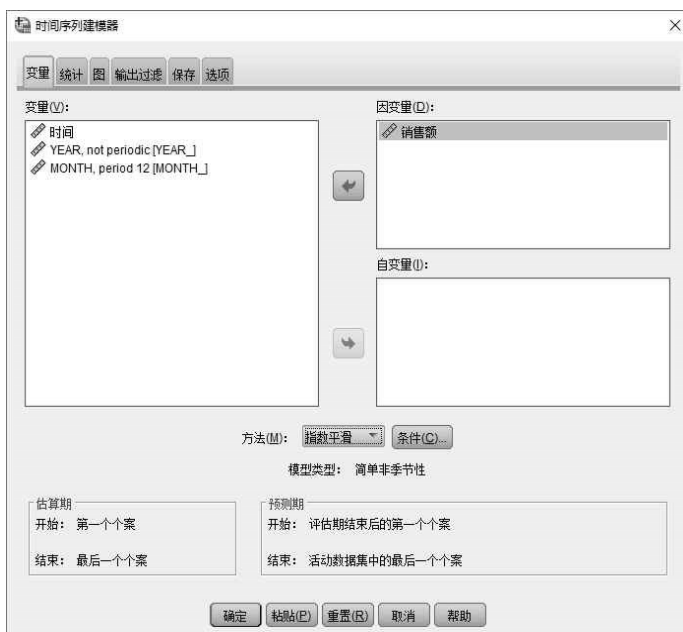


图 19-17 “时间序列建模器”对话框

指数平滑条件的设置, 选择具有“简单季节模型”(Simple Seasonal)选项, “因变量转换 (Dependent Variable Transformation)”变量框中选择“无 (None)”选项。



图 19-18 “指数平滑条件 (Criteria)” 的设置

单击图 19-17 中的“图 (Plots)” 标签，进行图形绘制参数设置，如图 19-19 所示。



图 19-19 “图表 (Plots) 设置” 对话框

选择“选项 (Options)” 标签，弹出如图 19-20 所示的对话框，设置情况如下，在“日期 (Date)” 选项栏中填入年 (Year) 为 2008，月 (Month) 为 1。

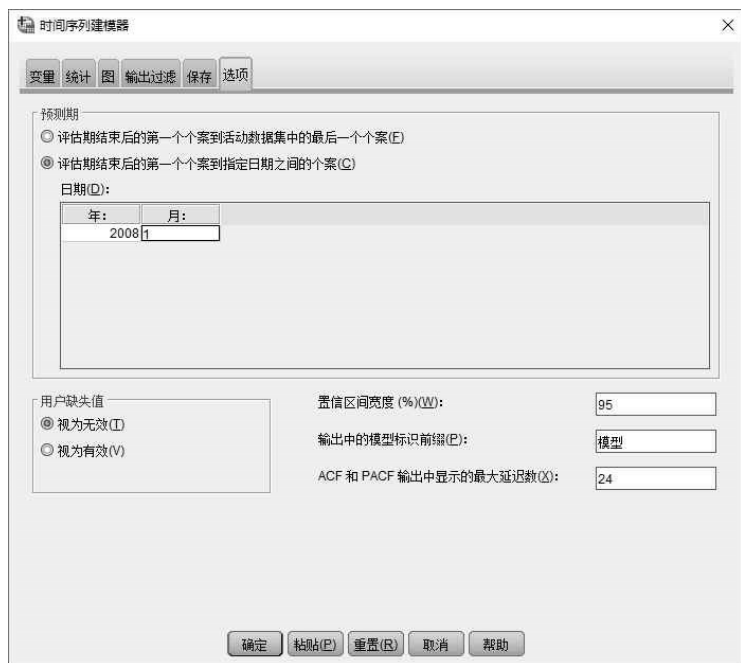


图 19-20 “选项 (Options) 设置”对话框

(2) 结果分析

设置完成以后,则单击“确定”按钮,进行系统运行,在 SPSS 输出中输出结果。首先是模型的描述,如图 19-21 所示。

| 模型描述 | | | 模型类型 |
|-------|-----|------|-------|
| 模型 ID | 销售额 | 模型_1 | 简单季节性 |

图 19-21 模型描述

然后输出的是模型拟合情况,包括 R 方在内的各种拟合统计量,如图 19-22 所示。

| 拟合统计量 | 平均值 | SE | 最小值 | 最大值 | 百分位数 | | | | | | |
|----------|--------|----|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| | | | | | 5 | 10 | 25 | 50 | 75 | 90 | 95 |
| 平稳的 R 方 | -1.323 | . | -1.323 | -1.323 | -1.323 | -1.323 | -1.323 | -1.323 | -1.323 | -1.323 | -1.323 |
| R 方 | .899 | . | .899 | .899 | .899 | .899 | .899 | .899 | .899 | .899 | .899 |
| RMSE | 5.526 | . | 5.526 | 5.526 | 5.526 | 5.526 | 5.526 | 5.526 | 5.526 | 5.526 | 5.526 |
| MAPE | 4.088 | . | 4.088 | 4.088 | 4.088 | 4.088 | 4.088 | 4.088 | 4.088 | 4.088 | 4.088 |
| MaxAPE | 17.128 | . | 17.128 | 17.128 | 17.128 | 17.128 | 17.128 | 17.128 | 17.128 | 17.128 | 17.128 |
| MAE | 3.692 | . | 3.692 | 3.692 | 3.692 | 3.692 | 3.692 | 3.692 | 3.692 | 3.692 | 3.692 |
| MaxAE | 13.706 | . | 13.706 | 13.706 | 13.706 | 13.706 | 13.706 | 13.706 | 13.706 | 13.706 | 13.706 |
| 正态化的 BIC | 3.740 | . | 3.740 | 3.740 | 3.740 | 3.740 | 3.740 | 3.740 | 3.740 | 3.740 | 3.740 |

图 19-22 模型拟合结果

然后输出的是残差的相关函数序列图。如图 19-23 所示是关于自相关 (ACF) 和偏相关 (PACF) 的序列图, 可以看出并没有显著的趋势特征, 所以, 可以使用该模型进行预测分析。

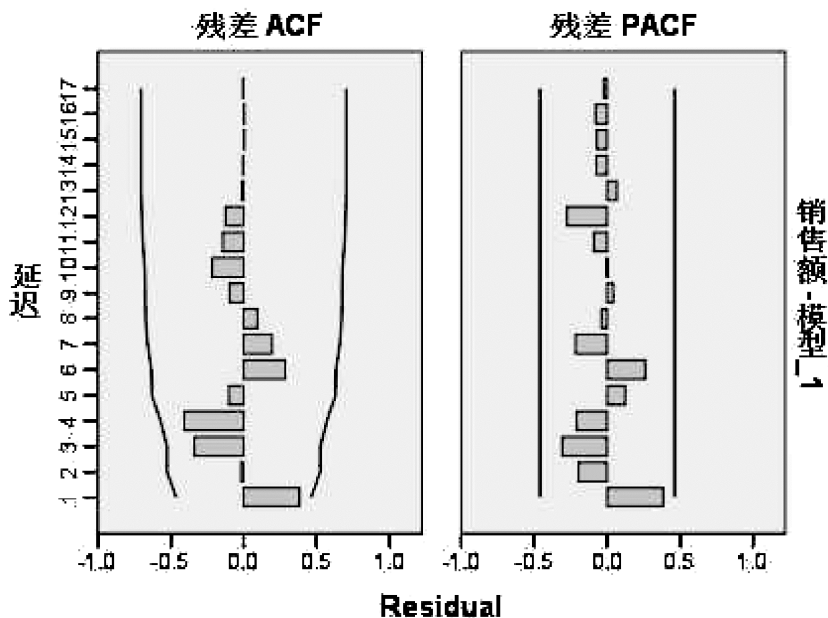


图 19-23 残差的相关函数序列图

最后是预测结果和拟合结果, 如图 19-24 所示, 包含置信区间的上限和下限。

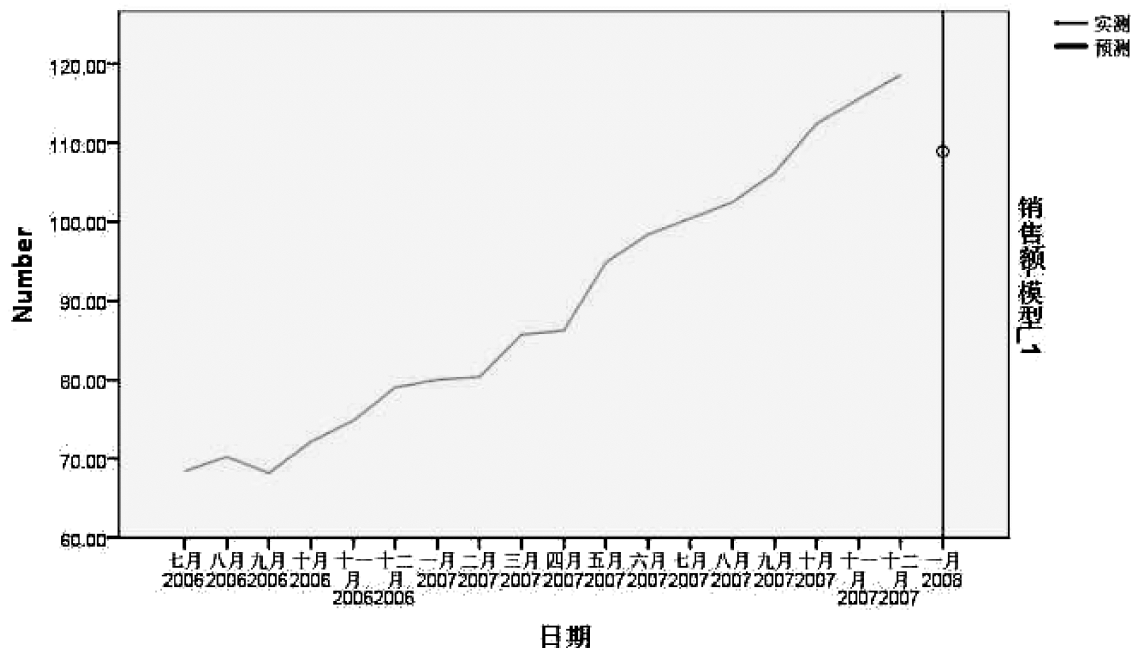


图 19-24 预测结果

19.4 ARIMA 模型

19.4.1 ARIMA 模型的基本原理

在 19.1 节, 已经对 ARIMA 模型进行了概述, 下面重点介绍 ARIMA 模型的识别和参数估计等内容。

1. ARIMA 模型的识别

采用 ARIMA 模型对现有的数据进行建模, 首要的问题是确定模型的阶数, 即相应的 p, q 值, 对于 ARIMA 模型的识别主要是通过序列的自相关函数和偏自相关函数进行的。

序列 y_t 的自相关函数度量了 y_t 与 y_{t-k} 之间的线性相关程度, 用 ρ_k 表示, 定义如下。

$$\rho_k = \frac{r_k}{r_0}$$

式中, $r_k = \text{cov}(y_t, y_{t-k})$; $r_0 = \text{cov}(y_t, y_t)$ 表示序列的方差。

自相关函数刻画的是 y_t 与 y_{t-k} 之间的线性相关程度, 而有时候 y_t 与 y_{t-k} 之间之所以存在相关关系, 可能是因为 y_t 和 y_{t-k} 分别与它们的中间部分 $y_{t-1}, y_{t-2}, \dots, y_{t-k+1}$ 之间存在关系, 如果在给定 $y_{t-1}, y_{t-2}, \dots, y_{t-k+1}$ 的前提下, 对 y_t 和 y_{t-k} 之间的条件相关关系进行刻画, 则要通过偏自相关函数 φ_{kk} 进行, 偏自相关函数可由下面的递推公式得到

$$\begin{cases} \varphi_{11} = \rho_1 \\ \varphi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \varphi_{k-1,j} \rho_{k-j}}{1 - \sum_{j=1}^{k-1} \varphi_{k-1,j} \rho_j} \\ \varphi_{k,j} = \varphi_{k-1,j} - \varphi_{kk} \varphi_{k-1,k-j}, \quad j=1, 2, \dots, k-1 \end{cases}$$

对于三类模型 AR, MA, ARIMA, 它们各自的自相关函数及偏自相关函数见表 19-2。

表 19-2 三类模型的相关函数

| 模型系数 | AR (p) | MA (q) | ARIMA (p, q) |
|-----------------------|--|------------------------------------|------------------|
| 自相关函数 ρ_k | 拖尾 | q 步截尾 ($\rho_k = 0, k > q$) | 拖尾 |
| 偏自相关函数 φ_{kk} | p 步截尾 ($\varphi_{kk} = 0, k > p$) | 拖尾 | 拖尾 |

这里的拖尾指模型自相关函数或偏自相关函数随着时滞 k 的增加呈现指数衰减并趋于零, 而截尾则是指模型的自相关函数或偏自相关函数在某步之后全部为零。序列的自相关函数和偏自相关函数所呈现出的这些性质可用于模型的识别。

(1) 基于自相关函数和偏自相关函数的定阶方法

理论上讲, 对于 $AR(p)$ 序列的偏自相关函数是 p 步截尾的, 但实际中所接触到的往往是来自序列的一组样本, 所计算的也只能是样本的偏自相关函数, 由于样本的随机性, 此时计算所得的样本偏自相关函数不可能是 p 步截尾的, 而是在零附近波动, 所以, 要考虑的是样本偏自相关函数的统计性质, 对于 $MA(q)$ 序列的样本自相关函数同样应该考虑其统计性质。关于样本自相关函数 $\hat{\rho}_k$ 的估计方法很多, 最常用的是如下的估计方法, 即

$$\begin{cases} \bar{y} = \frac{1}{n} \sum_{t=1}^n y_t \\ \hat{\gamma}_k = \frac{1}{n} \sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y}) \\ \hat{\rho}_k = \hat{\gamma}_k / \hat{\gamma}_0, \quad k = 0, 1, 2, \dots, n-1 \end{cases}$$

式中, \bar{y} 为样本均值; $\hat{\gamma}_k$ 为样本自协方差函数。

(2) 利用信息准则法定阶

信息准则法在模型选择中起到很重要的作用, 关于定阶问题, 实际上也是模型选择问题, 这里给出两种准则。

1) AIC

AIC 准则 (Akaike's Information Criterion, 赤池信息准则), 是由 Akaike 在 1973 年提出的, 该准则既考虑拟合模型对数据的接近程度, 也考虑模型中所含待定参数的个数。关于 $ARIMA(p, q)$, 对其定义的 AIC 函数如下:

$$AIC(p, q) = n \ln(\hat{\sigma}^2) + 2(p + q)$$

其中 $\hat{\sigma}^2$ 是拟合 $ARIMA(p, q)$ 模型时残差的方差, 它是 (p, q) 的函数。如果模型中含有常数项, 则 $p + q$ 被 $p + q + 1$ 代替。AIC 定阶的方法就是选择 $AIC(p, q)$ 最小的 (p, q) 作为相应的模型阶数。

2) BIC

Akaike 在 1976 年改进了 AIC 准则, 提出 BIC 准则。这样避免了在大样本情况下, AIC 准则在选择阶数收敛性不好的缺点。关于 $ARIMA(p, q)$, 对其定义的 BIC 函数如下:

$$AIC(p, q) = n \ln(\hat{\sigma}^2) + 2(p + q) \ln n$$

BIC 定阶的方法就是选择 $AIC(p, q)$ 最小的 (p, q) 作为相应的模型阶数。

利用 AIC 准则和 BIC 准则确定出来的 $ARIMA$ 模型可能不一致, 一般说来, 用 BIC 准则选择出来的 $ARIMA$ 模型的阶数较 AIC 准则选择的阶数低。

2. 模型参数的估计

模型的阶数确定之后, 就可以估计模型了。主要有三种估计方法: 矩估计, 极大似然估计和最小二乘估计。最小二乘估计和极大似然估计的精度较高, 因而一般称为模型参数的精估计。最小二乘估计在一般的数理统计教材中都有全面的介绍, 本书不再重述。而极大似然估计计算方法较为复杂, 最后求解的方程皆为非线性方程, 很难求解, 所以实际中采用数值算法。思路是任意给出参数的一组数值, 初步估计得到的结果, 计算出一个似然函数值; 然后, 根据一定的法则, 再给出参数的一组数值, 又计算出一个似然函数值; 依

此类推, 比较似然函数值, 选择使似然函数值最大的那组参数。本节主要介绍矩估计法, 以 AR 模型为例说明方法, MA 和 ARIMA 模型思路相同, 只是步骤更复杂一些, MA 和 ARIMA 的矩估计再详细介绍。

下面是一个零均值的 AR(p)模型, 即

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

需要估计的参数是 $\phi_1, \phi_2, \cdots, \phi_p$ 。

在模型两边同乘以 $y_{t-j}, j > 0$, 可得

$$y_t y_{t-j} = \phi_1 y_{t-1} y_{t-j} + \phi_2 y_{t-2} y_{t-j} + \cdots + \phi_p y_{t-p} y_{t-j} + \varepsilon_t y_{t-j}$$

两边取期望, 得

$$E y_t y_{t-j} = \phi_1 E y_{t-1} y_{t-j} + \phi_2 E y_{t-2} y_{t-j} + \cdots + \phi_p E y_{t-p} y_{t-j} + E \varepsilon_t y_{t-j}$$

由于 ε_t 与 $y_{t-j} (j > 0)$ 不相关, 所以 $E \varepsilon_t y_{t-j} = 0$, 因此

$$r_j = \phi_1 r_{j-1} + \phi_2 r_{j-2} + \cdots + \phi_p r_{j-p}, \quad j > 0$$

其中, r_j 是序列 $\{y_t\}$ 的自协方差函数, 易知序列的自相关函数 ρ_j 也满足上述关系式, 即

$$\rho_j = \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2} + \cdots + \phi_p \rho_{j-p}, \quad j = 1, 2, 3, \cdots$$

把自相关函数展成 p 个方程, 即

$$\begin{cases} \rho_1 = \phi_1 \rho_0 + \phi_2 \rho_1 + \cdots + \phi_p \rho_{p-1} \\ \rho_2 = \phi_1 \rho_1 + \phi_2 \rho_0 + \cdots + \phi_p \rho_{p-2} \\ \vdots \\ \rho_p = \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + \cdots + \phi_p \rho_0 \end{cases}$$

上述 p 个方程, 表示了平稳序列的自相关函数与模型未知参数的关系, 称为 Yule-Walker 方程。

自相关函数可以用样本自相关函数代替, 所以, 此时的 Yule-Walker 方程只有 p 个未知数, 解方程可以得到 $\phi_1, \phi_2, \cdots, \phi_p$ 的估计值, 用矩阵表示如下, 即

$$\begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_p \end{bmatrix} = \begin{bmatrix} 1 & \hat{\rho}_1 & \cdots & \hat{\rho}_{p-1} \\ \hat{\rho}_1 & 1 & \cdots & \hat{\rho}_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_{p-1} & \hat{\rho}_{p-2} & \cdots & 1 \end{bmatrix}^{-1} \begin{bmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \vdots \\ \hat{\rho}_p \end{bmatrix}$$

对于二阶自回归模型 AR(1), 根据上述结果可知

$$\hat{\phi}_1 = \hat{\rho}_1$$

对于二阶自回归模型 AR(2), 根据上述结果可知

$$\begin{cases} \hat{\phi}_1 = \frac{\hat{\rho}_1(1 - \hat{\rho}_2)}{1 - \hat{\rho}_1^2} \\ \hat{\phi}_2 = \frac{\hat{\rho}_2 - \hat{\rho}_1^2}{1 - \hat{\rho}_1^2} \end{cases}$$

样本自相关函数和自协方差函数除了定阶外，还可以用来估计。矩估计也称为初估计，矩估计方法简单但精度不高。

19.4.2 ARIMA 模型分析过程的参数设置

选择图 19-2 中的“方法 (Method)”下拉菜单，选择 ARIMA 模型选项，然后单击“条件 (Criteria)”按钮，则弹出如图 19-25 所示的对话框，此对话框可以设置 ARIMA 模型的参数。

1. 模型 (Model) 设置

此选项框用于设置模型参数。

ARIMA 阶数 (ARIMA Orders) 选项：用于指定不同成分的阶数，以确定模型的结构。

- 非季节性 (Nonseasonal)：从上到下依次为 SARIMA(p,d,q)*(sp,sd,sq)模型的 p, d, q。
- 季节性 (Seasonal)：从上到下依次为 SARIMA(p,d,q)*(sp,sd,sq)模型的 sp, sd, sq。
- 当前周期长度 (Current Periodicity)：显示数据周期。

转换 (Dependent Variable Transformation)：用于指定因变量的变换方法。

- 无 (None)：不作变换。
- 平方根 (Square Root)
- 自然对数 (Natural Log)

在模型中包括常数 (Include Constant in Model)：表示在 ARIMA 模型中包含常数项。

2. 离群值 (Outliers) 设置

单击图 19-25 中的“离群值 (Outliers)”标签，则弹出如图 19-26 所示的对话框，此对话框用于设置异常值检测选项。

不检测离群值，也不为其建模 (Do not Detect Outlier Or Model them) 栏：表示不进行异常值的检测。

自动检测离群值 (Detect Outliers Automatically) 栏：此栏指定自动检测异常值方法，有如下几个选择项。

- 加法 (Additive)：表示只影响单个观测记录的异常值。
- 水平变动 (Level Shift)：由于数据水平移动而引起的异常值。
- 革新 (Innovational)：由于噪声变动形成的异常值。
- 瞬态 (Transient)：对后续观测的影响程度。
- 季节加性 (Seasonal Additive)：周期性影响某些时刻的异常值，而且其影响程度对不同时刻的观测是相同的。
- 局部趋势 (Local Trend)：局部的线性异常值。
- 加性修补 (Additive Patch)：表示两个或者多个连续出现的可加 (Additive) 类型的异常值。

将特定的时间点作为离群值进行建模 (Model specific time points as outliers) 栏：用来设置特定时刻的数据位异常值，此栏选中后可以在离群值定义 (Outlier Definition) 下的二维表格中每行指定一个特定的异常数据，第一列输入时间点，第二列类型 (Type) 从下拉菜单中选择异常点的类型。





图 19-25 “ARIMA 模型设置”对话框



图 19-26 “离群值 (Outliers) 设置”对话框

19.4.3 实例分析

-  **结果文件** —— 附带光盘 “PROGRAM\CH19\实例 19-2” 文件夹
-  **动画演示** —— 附带光盘 “AVI\实例 19-2.avi” 文件

本案例利用 ARIMA 过程对上海证券交易所综合指数收益率序列进行模拟, 以期对股市走势有深刻的认识。上证综指的数据为 2005 年 1 月 4 日到 2009 年 7 月 31 日的日交易数据集 ARIMA2.sav, 此数据集在 SPSS 中的数据格式如图 19-27 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|-----|----|----|------|----|---|----|----|----|----|----|
| 1 | 时间 | 数字 | 11 | 0 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 2 | 收盘价 | 数字 | 11 | 3 | | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 3 | 收益率 | 数字 | 11 | 9 | | 无 | 无 | 11 | 右 | 标度 | 输入 |

图 19-27 ARIMA2.sav 数据格式

下面就对此数据集进行时间序列预测分析。

1. 参数设置

首先定义时间变量, 选择菜单“数据 (Data) 定义日期和时间 (Define Dates)”, 打开如图 19-28 所示对话框, 然后选择变量“天 (Days)”, 在其右边对话框中填入 1, 然后单击“确定 (OK)”按钮返回主界面, 如图 19-28 所示。

然后选择菜单“分析 (Analyze) 时间序列预测 (Forecast) 创建传统模型 (Create Models)”, 弹出如图 19-29 所示对话框, 选入变量收盘价到“因变量 (Dependent

Variables)”变量框中,然后选择“方法(Method)”下拉菜单的“专家建模器(Expert Modeler)”选项。



图 19-28 “定义时间和日期(Define Dates)”对话框



单击“条件(Criteria)”按钮,弹出如图 19-30 所示对话框,用于设置使用模型。选择“仅限 ARIMA 模型(ARIMA Models Only)”选项,然后单击“继续(Continue)”按钮返回主界面。

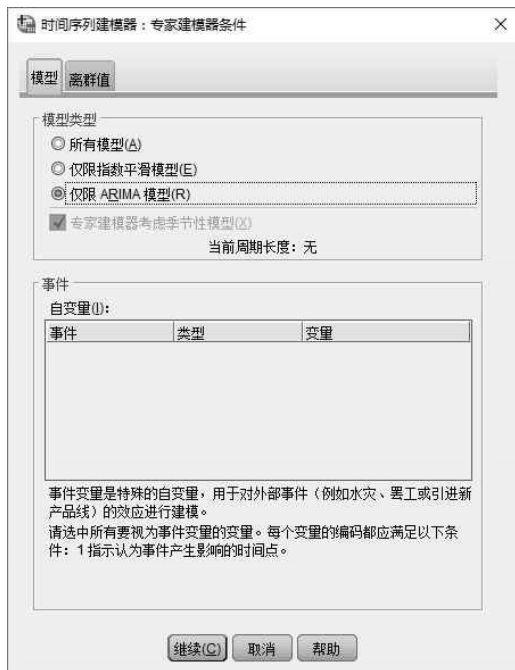


图 19-30 “条件(Criteria)设置”对话框

其他方面的设置,单击图 19-29 中的“图(Plots)”标签,弹出如图 19-31 所示对话框,用于设置绘制图形的各种参数,设置情况如图 19-31 所示。



图 19-31 “图 (Plots) 设置”对话框

2. 结果分析

单击主界面时间序列预测建模器 (Time Series Modeler) 中的“确定 (OK)”按钮, 则系统进行时间序列预测分析, 首先是模型的描述, 如图 19-32 所示。给出了 ARIMA 模型中的最佳参数, 其中 p 、 d 、 q 分别是 2、1、2。

| 模型描述 | | | |
|-------|-----|------|--------------|
| | | | 模型类型 |
| 模型 ID | 收盘价 | 模型_1 | ARIMA(2,1,2) |

图 19-32 模型的描述

然后是模型拟合结果, 如图 19-33 所示。包括各种拟合优度的检验统计量。

| 拟合统计 | 平均值 | SE | 最 小 值 | 最 大 值 | 百 分 位 数 | | | | | | |
|----------|---------|----|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| | | | | | 5 | 10 | 25 | 50 | 75 | 90 | 95 |
| 平稳的 R 方 | .003 | . | .003 | .003 | .003 | .003 | .003 | .003 | .003 | .003 | .003 |
| R 方 | .998 | . | .998 | .998 | .998 | .998 | .998 | .998 | .998 | .998 | .998 |
| RMSE | 64.160 | . | 64.160 | 64.160 | 64.160 | 64.160 | 64.160 | 64.160 | 64.160 | 64.160 | 64.160 |
| MAPE | 1.474 | . | 1.474 | 1.474 | 1.474 | 1.474 | 1.474 | 1.474 | 1.474 | 1.474 | 1.474 |
| MaxAPE | 9.732 | . | 9.732 | 9.732 | 9.732 | 9.732 | 9.732 | 9.732 | 9.732 | 9.732 | 9.732 |
| MAE | 40.642 | . | 40.642 | 40.642 | 40.642 | 40.642 | 40.642 | 40.642 | 40.642 | 40.642 | 40.642 |
| MaxAE | 349.964 | . | 349.964 | 349.964 | 349.964 | 349.964 | 349.964 | 349.964 | 349.964 | 349.964 | 349.964 |
| 正态化的 BIC | 8.335 | . | 8.335 | 8.335 | 8.335 | 8.335 | 8.335 | 8.335 | 8.335 | 8.335 | 8.335 |

图 19-33 模型拟合结果

下面是 ARIMA (2,1,2) 模型的参数输出,如图 19-33 所示的输出结果,图 19-34 是 ARIMA 模型残差的相关函数图形,如图 19-35 所示是关于残差序列的自相关 (ACF) 图形和偏相关 (PACF) 图形。

| 模型统计 | | | | | | |
|----------|-------|---------|--------|-------------|------|------|
| 模型 | 预测变量数 | 模型拟合度统计 | | 杨-博克斯 Q(18) | | 离群值数 |
| | | 平稳 R 方 | 统计 | DF | 显著性 | |
| 收盘价-模型_1 | 0 | .003 | 33.177 | 16 | .007 | 0 |

图 19-34 ARIMA 模型参数

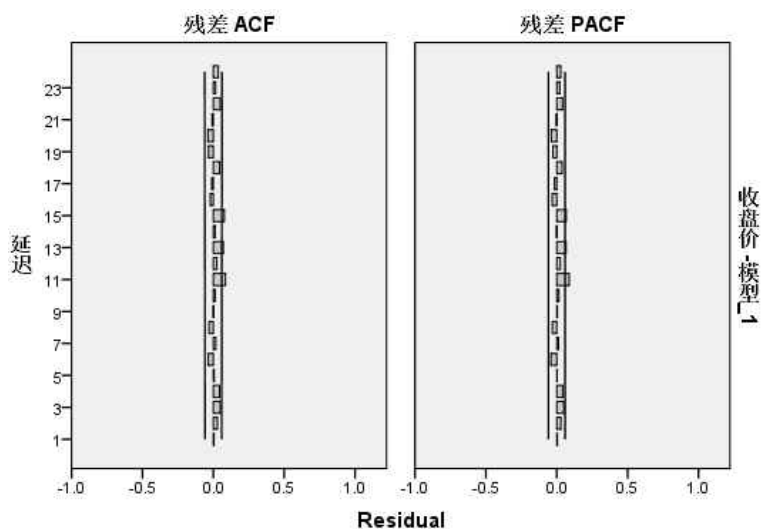


图 19-35 相关函数图形

最后输出的是 ARIMA 模型预测的结果和拟合结果图形,如图 19-36 所示包含置信区间的上限和下限。另外,关于 ARIMA 模型的预测结果会自动保存在原始数据集中,方便用户查询,在此不再赘述。

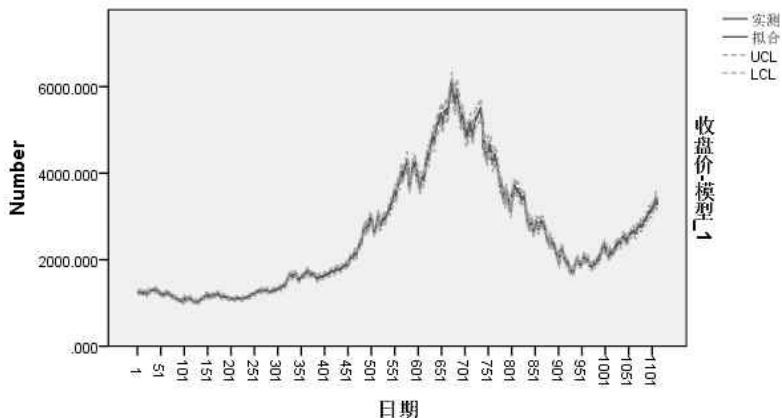


图 19-36 ARIAM 模型拟合结果

19.5 季节性分解模型

季节变动是指现象在一定时期内形成的有规律的周期性变动,这种变动各年强度大体相同且重复出现。测定季节变动的目的在于了解现象季节变动的规律,能进行预测。

季节变动的测定主要是计算一系列季节指数,又称季节比率,其设计思想是,以总平均水平为对照物,用各季节的平均数与之比较,来反映季节变动高低程度。季节指数是各季(月)平均数与全时期总平均数的比率,它由一系列数值组成,个数由资料的时间间隔决定,且季节指数之和也与所掌握资料有关。如掌握资料为月份资料,则有12个季节指数,季节指数之和为1200%,如为季度资料,则有4个季节指数,季节指数之和为400%。

下面从时间序列是否包含长期趋势方面来介绍测定季节变动的方法。

(1) 不包含长期趋势的时间序列

若时间序列中不包含长期趋势和循环变动,则直接利用原序列进行同期平均和总平均,消除不规则变动,计算出季节指数,常用按季(月)平均法,基本步骤如下。

计算同月(或同季)的平均数。

计算全部数据的总月(总季)平均数。

计算季节指数(S),即 $S = \frac{\text{各月平均数}}{\text{总平均数}}$ 。

(2) 包含长期趋势的时间序列

当时间序列包含长期趋势和循环变动时,用按季平均法计算季节指数就不够准确,应采用趋势剔除法。假定时间序列各影响因素以乘法模型形式存在,趋势剔除法的基本步骤如下。

用移动平均法、趋势线法等方法消除季节变动(S)和不规则(I)变动,计算出长期趋势和循环变动值($T \times C$)。

再从乘法模型中剔除($T \times C$),从而得到不存在长期趋势的($S \times I$),即

$$S \times I = \frac{Y}{T \times C}$$

再用按季(月)平均法消除 I ,得到季节指数。

19.5.1 季节性分解模型分析过程的参数设置

选择菜单“分析(Analyze) 时间序列预测(Forecast) 季节性分解(Seasonal Decomposition)”,弹出如图19-37所示的对话框,此界面用来设置季节性分解模型的各个参数。

变量(Variable(s)):此栏用于选入进行季节性分解的原始序列变量。

模型类型(Model Type):此栏用于指定季节性分解的模型类型。

- 乘性(Multiplicative)
- 加性(Additive)

移动平均值权重(Moving Average Weight):用于指定计算移动平均时的权重。

- 所有点相等(All Points Equal)
- 端点按0.5加权(Endpoints Weighted by 0.5)

当前周期长度 (Current Periodicity): 显示当前数据的周期。

显示个案列表。

保存 (Save) 设置。

单击“保存 (Save)”按钮, 则弹出如图 19-38 所示的对话框。

- 添加至文件 (Add to file): 作为永久新增变量加入到当前数据集里。
- 替换现有 (Replace existing): 作为临时新增变量加入到当前数据集里, 新的模型输出将覆盖旧模型保存的变量。
- 不要创建 (Do not create): 不在当前数据集保存模型结果。

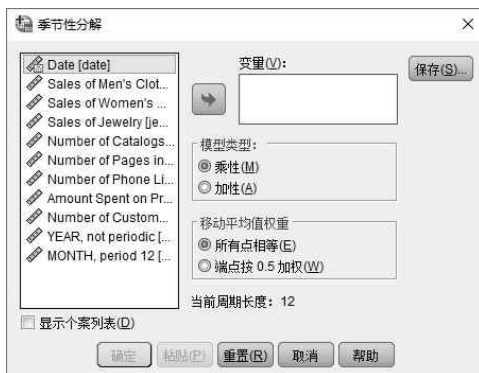


图 19-37 “季节性分解 (Seasonal Decomposition) 设置”对话框 图 19-38 “保存 (Save)”对话框

19.5.2 实例分析

结果文件——附带光盘“PROGRAM\CH19\实例 19-3”文件夹

动画演示——附带光盘“AVI\实例 19-3.avi”文件

下面就利用 SPSS 软件对数据集进行季节性分解模型的分析, 所用数据集为 SPSS 自带的 catalog.sav, 此数据集为 catalog 公司的服装销售数据, 数据集包含变量 date、men、women、jewel、mail、page、phone、print、service, 数据集从 1989 年 1 月 1 日到 1998 年 12 月 1 日共 10 年的每月销售数据综合, 下面要对此数据集进行趋势分析。数据集格式如图 19-39 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|-------|----|----|------|--------------------|----|---|----|----|----|
| 1 | date | 日期 | 11 | 0 | Date | 无 | 8 | 右 | 标度 | 输入 |
| 2 | men | 数字 | 11 | 2 | Sales of Men's ... | 无 | 8 | 右 | 标度 | 输入 |
| 3 | women | 数字 | 11 | 2 | Sales of Wome... | 无 | 8 | 右 | 标度 | 输入 |
| 4 | jewel | 数字 | 13 | 12 | Sales of Jewelr... | 无 | 8 | 右 | 标度 | 输入 |
| 5 | mail | 数字 | 11 | 0 | Number of Cata... | 无 | 8 | 右 | 标度 | 输入 |
| 6 | page | 数字 | 11 | 0 | Number of Pag... | 无 | 8 | 右 | 标度 | 输入 |
| 7 | phone | 数字 | 11 | 0 | Number of Pho... | 无 | 8 | 右 | 标度 | 输入 |

图 19-39 catalog.sav 数据集格式

首先要判断是否具有季节周期性,以便消除季节变量的影响。应用季节性分解(Seasonal Decomposition)过程,首先分析男性服装的数据分布情况。选择菜单“分析(Analyze) 时间序列预测(Forecast) 序列图(Sequence Charts)”,则弹出如图 19-40 所示对话框,选择变量 Sales of Men's Clothing 到“变量(Variables)”选项栏中,选择变量 Date 到“时间轴标签(Time Axis Labels)”选项栏中,然后单击“确定”按钮,则系统进行绘制图形操作,结果如图 19-41 所示。



图 19-40 “序列图(Sequence Charts)设置”对话框

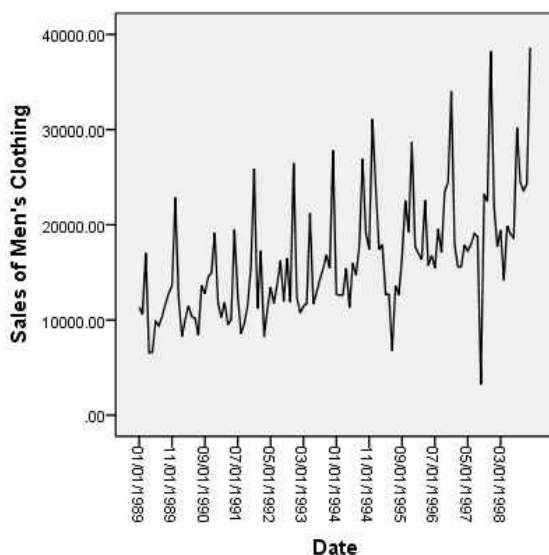


图 19-41 时间序列预测图形

从图 19-41 中可以明显看出,数据具有季节周期性,所以,进行趋势分析时要消除季节性因素的影响。下面检验时间序列的自相关和偏相关系数。选择菜单“分析(Analyze) 时间序列预测(Forecast) 自相关(Autocorrelations)”,打开如图 19-42 所示的对话框,把变量 Sales of Men's Clothing 选入“变量(Variables)”选项框中,然后单击“确定(OK)”按钮,则输出分析结果。



图 19-42 “自相关和偏相关系数设置”对话框

图 19-43 是自相关系数分析结果,置信区间最大点是坐标 12 所对应的数值。图 19-44 是偏相关系数分析结果图形,还是于坐标 12 处的最大。



图 19-43 自相关系数分析结果



图 19-44 偏相关系数分析结果

下面根据上述分析来确定平均周期，选择菜单“数据（Data） 定义日期和时间（Define Dates）”，弹出如图 19-45 所示对话框。在“个案是（Cases Are）”选项栏中选中年、月（Years, Months）变量，然后在图 19-45 右边的“年（Year）”和“月（Month）”选项栏中填入 1989 和 1，然后单击“确定（OK）”按钮进行分析。

然后，进行季节性分解（Seasonal Decomposition）过程分析，选择菜单“分析（Analyze） 时间序列预测（Forecast） 季节性分解（Seasonal Decomposition）”，则弹出如图 19-46 所示对话框，选入变量 Sales of Men's Clothing 到“变量（Variables）”选项框，并选中“乘性（Multiplicative）”选项，然后单击“确定”按钮，系统进行分析，输出大量结果。其中 SAF 为季节调整因子；SAS 为季节调整后的序列；STC 为进行趋势光滑操作后的结果，包括趋势项、周期项；ERR 为误差项。

季节性分解的结果如图 19-47 所示，包括上述的几项分析结果。



图 19-45 时间周期确定

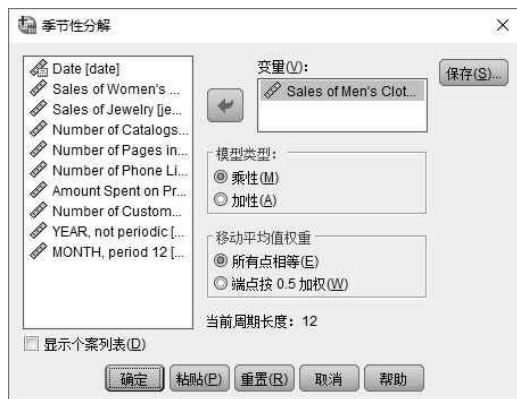


图 19-46 季节性分解（Seasonal Decomposition）过程分析

下面绘制季节调整后的时间序列图形，选择菜单“分析（Analyze） 时间序列预测（Forecast） 序列图（Sequence Charts）”，则弹出如图 19-48 所示对话框，把变量 SAS_1 选入“变量（Variables）”选项栏中，然后单击“确定”按钮进行分析。



Figure 19-47 displays the SPSS Data Editor window showing the results of a seasonal analysis. The data is organized into columns labeled ERR_1, SAS_1, SAF_1, and STC_1, with rows representing different time periods. The variable 'SEASON' is selected, and the results are displayed in a table format.

| | ERR_1 | SAS_1 | SAF_1 | STC_1 | 变量 |
|----|---------|-------------|---------|-------------|----|
| 1 | 76723 | 11932.95470 | 95181 | 15553.39388 | |
| 2 | 84698 | 12550.90257 | 84503 | 14818.39279 | |
| 3 | 1.49616 | 19971.32110 | 85115 | 13348.39061 | |
| 4 | 67845 | 7727.04200 | 84945 | 11389.26250 | |
| 5 | 75375 | 7732.70835 | 85451 | 10259.00632 | |
| 6 | 1.13411 | 11363.27501 | 86586 | 10019.52919 | |
| 7 | 1.01056 | 10980.97674 | 85587 | 10866.21741 | |
| 8 | 96073 | 10931.52550 | 95097 | 11378.30895 | |
| 9 | 1.07224 | 12530.71712 | 93076 | 11686.48707 | |
| 10 | 94416 | 11223.54133 | 1.14093 | 11887.34416 | |
| 11 | 1.00343 | 12295.12188 | 1.10908 | 12253.06026 | |
| 12 | 1.04457 | 12732.26995 | 1.79457 | 12189.00936 | |

图 19-47 季节性分析结果



图 19-48 “序列图 (Sequence Charts) 设置”对话框

单击“确定”按钮后,则输出绘制结果,如图 19-49 所示。调整后的时间序列给出了一个非常清晰的向上趋势。

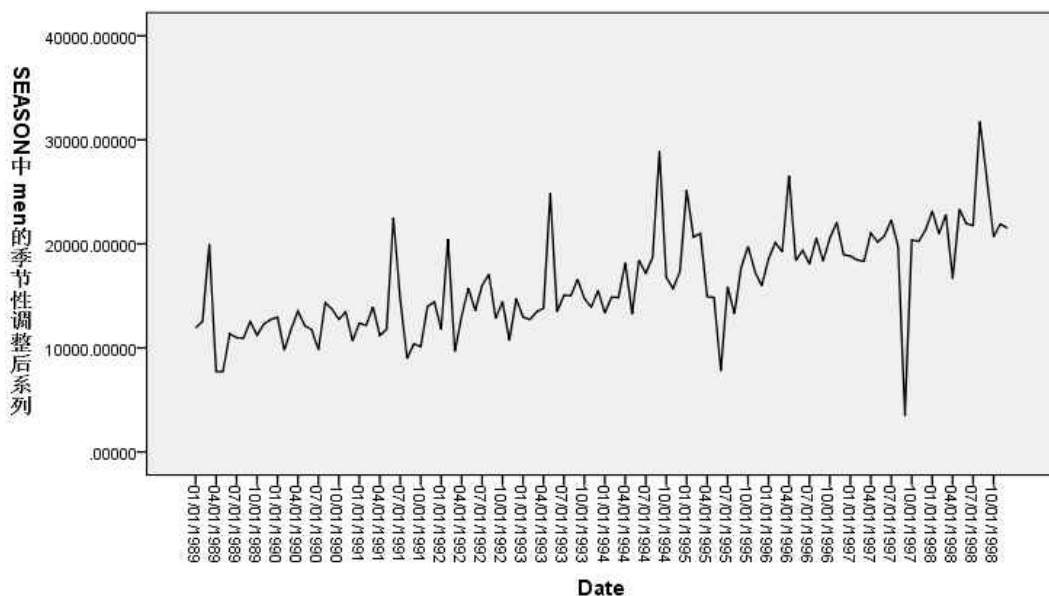
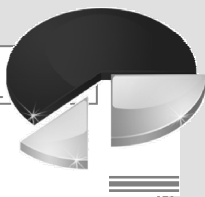


图 19-49 调整后的时间序列预测图形



第 20 章 缺失值分析

数据缺失在许多研究领域都是一个复杂的问题。对数据挖掘来说，空值的存在，造成了以下影响：第一，系统丢失了大量的有用信息；第二，系统中所表现出的不确定性更加显著，系统中蕴涵的确定性成分更难把握；第三，包含空值的数据会使挖掘过程陷入混乱，导致不可靠的输出。

本章将介绍利用 SPSS 软件系统来进行缺失值的处理方法，以便数据分析的顺利展开。



本讲内容

- 缺失值概述
- SPSS 缺失值分析设置
- 缺失值实例分析

20.1 缺失值理论概述

众所周知，在如收入、交通事故等问题的调查研究中，存在大量未回答的问题，以下是一些例子。

- 在一项消费者经济状况调查中，28%的受访者没有回答收入情况。
- 在一次人口调查中，20%的人没有回答收入情况，高收入者的回答率比中等收入者要低。
- 在严重交通事故报告中，如是否使用安全带和酒精浓度等关键问题在很多个案中都没有记录。

造成数据缺失的原因主要如下。

有些信息暂时无法获取，或者获取信息的代价太大。

有些信息是被遗漏的。可能是因为输入时认为不重要、忘记填写了或对数据理解错误而遗漏，也可能是由于数据采集设备的故障、存储介质的故障、传输媒体的故障、一些人为因素等原因而丢失了。

有些对象的某个或某些属性是不可用的。也就是说，对于这个对象来说，该属性值是不存在的，如一个未婚者的配偶姓名、一个儿童的固定收入状况等。

概括大部分缺失值的情况，缺失值经常在下列一些情况中出现。

- 拒绝回答问题。
- 没有答案。
- 调查研究中的损耗。
- 从多个数据源中合并数据。

缺失值会表现为以下问题。

- 有缺失值的个案系统地不同于完整的个案。
- 有缺失值的个案表明信息不完整。
- 标准统计方法只接受完整数据。

以上问题意味着如下方面。

- 偏向：分析结果可能会有偏差。
- 无效：较少的有效个案导致估计精度下降。

20.1.1 数据缺失方式

数据缺失的大致方式可以分为如下几种。

1. 数据完全随机缺失 (Missing Completely at Random, MCAR)

数据完全随机缺失表示缺失和变量的取值无关。例如，假设你在研究年龄和收入，如果缺失和年龄或收入数值无关，则缺失值方式为 MCAR。要评估 MCAR 是否为合适的假设，你可以用比较回答者和未回答者的分布来评估观察数据。也可以使用单变量 t 检验或利特尔 MCAR 多变量检验来进行更正规的评估。如果 MCAR 假设为真，可以使用列表删除（按列表（Listwise）deletion）（完整个案分析），无须担心估计偏差，尽管可能会丧失一些有效性。如果 MCAR 不成立，列表删除、均值置换等逼近方法就可能不是好的选择。

2. 随时缺失 (Missing at Random, MAR)

如果数据不为 MCAR，可以考虑评估回答者和未回答者的特性差异是否能够用同时测度回答者和未回答者的变量来理解。这就引出了随时缺失的概念，其中缺失分布中调查变量只依赖于数据组数中有记录的变量。继续上面的例子，考虑到年龄全部被观察，而且收入有时有缺失。这样，如果收入缺失值仅依赖于年龄，缺失值就为 MAR。如果收入缺失值依赖于收入值，则既不是 MCAR，也不是 MAR。

区别 MCAR 和 MAR 的含义在于，由于 MCAR 通常实际上很难遇到，应该在进行调查之前就考虑哪些重要变量可能会有无效的未回答，还要尽量在调查中包括共变量，以便用这些变量来估算缺失值。

20.1.2 缺失值处理方法

数据缺失在许多研究领域都是一个复杂的问题。对数据挖掘来说，空值的存在，造成了以下影响：第一，系统丢失了大量的有用信息；第二，系统中所表现出的不确定性更加显著，系统中蕴涵的确定性成分更难把握；第三，包含空值的数据会使挖掘过程陷入混

乱，导致不可靠地输出。

缺失值的处理一般主要有两种方式：一是删除对应的记录；二是进行插值处理。

1. 随机插值

根据缺失值的各种可能情况，等概率地进行插值。

例如，在上例中，“张三”的性别有两种可能性，一是“男”，二是“女”，可以简单地掷一枚硬币，如果正面朝上，则赋值为“男”，如果反面朝上，则赋值为“女”。

2. 依概率插值

随机插值是假设一个变量取各种值的可能性是相等的，但有些情况下，可以事先知道一个变量取各种值的概率，例如，我们知道在上述的单位中，女性占的比例是 75%，男性的比例是 25%，则在对“张三”的性别进行赋值时，不是按 50% 概率赋为“女”，而是按 75% 概率赋为“女”。

3. 就近插值

就近插值是指根据缺失记录附近的其他记录的情况对缺失值进行插值，例如，在上例中，“张三”的性别出现缺失，此时可以用其邻近的“李四”的性别数据替代“张三”的性别数据，由于“李四”的性别为“女”，所以将“张三”的性别也赋为“女”。

就近插值是依概率插值的一种简化处理，设想在整个单位的职工中，女性占的比例是 75%，则在一般情况下，与张三邻近的记录性别为“女”的概率也应当为 75%，就近插值实际上就是依概率插值。

使用就近插值时，需要对抽样过程进行必要的了解，如果抽样时性别有交叉的情况，例如，经常是调查完一名男性后就调查一名女性，则使用就近插值就会出现较多的错误。

4. 分类插值

依概率插值是将记录置于总体的背景上进行插值，没有充分利用记录的其他信息。如果在记录的其他信息中有某些项目与缺失项目存在相关性，则可以根据这些辅助信息对总体进行分类，在每一类内部进行插值处理。

值得注意的是，这里所说的缺失值，不仅包括数据库中的 NULL 值，也包括用于表示数值缺失的特殊数值（例如，在系统中用 -999 来表示数值不存在）。如果仅有数据库的数据模型，而缺乏相关说明，常常需要花费更多的精力来发现这些数值的特殊含义。而如果漠视这些数值的特殊性，直接拿来挖掘，那么很可能会得到错误的结论。

数据挖掘算法本身更致力于避免数据过分适合所建的模型，这一特性使得它难以通过自身的算法去很好地处理不完整数据。因此，空缺的数据需要通过专门的方法进行推导、填充等，以减少数据挖掘算法与实际应用之间的差距。

包括上面的数据插值方法，处理不完备数据集的方法主要有以下三大类。

1. 删除元组

也就是将存在遗漏信息属性值的对象（元组、记录）删除，从而得到一个完备的信息表。这种方法简单易行，在对象有多个属性缺失值、被删除的含缺失值的对象与信息表中

的数据量相比非常小的情况下是非常有效的,类标号(假设是分类任务)缺少时通常使用。然而,这种方法却有很大的局限性。它是以减少历史数据来换取信息的完备,会造成资源的大量浪费,丢弃了大量隐藏在这些对象中的信息。在信息表中本来包含的对象很少的情况下,删除少量对象就足以严重影响到信息表信息的客观性和结果的正确性;当每个属性空值的百分比变化很大时,它的性能非常差。因此,当遗漏数据所占比例较大,特别当遗漏数据非随机分布时,这种方法可能导致数据发生偏离,从而引出错误的结论。

2. 数据补齐

这类方法是用一定的值去填充空值,从而使信息表完备化。通常基于统计学原理,根据决策表中其余对象取值的分布情况来对一个空值进行填充,如用其余属性的平均值来进行补充等。数据挖掘中常用以下几种补齐方法。

(1) 人工填写 (Filling Manually)

由于最了解数据的还是用户自己,因此,这个方法产生数据偏离最小,可能是填充效果最好的一种。然而一般来说,该方法很费时,当数据规模很大、空值很多的时候,该方法是不可行的。

(2) 特殊值填充 (Treating Missing Attribute Values as Special Values)

将空值作为一种特殊的属性值来处理,它不同于其他的任何属性值。如果所有的空值都用“unknown”填充,这样将形成另一个有趣的概念,可能导致严重的数据偏离,一般不推荐使用。

(3) 平均值填充 (Mean/Mode Completer)

将信息表中的属性分为数值属性和非数值属性来分别进行处理。如果空值是数值型的,就根据该属性在其他所有对象的取值的平均值来填充该缺失的属性值;如果空值是非数值型的,就根据统计学中的众数原理,用该属性在其他所有对象的取值次数最多的值(出现频率最高的值)来补齐该缺失的属性值。另外有一种与其相似的方法叫条件平均值填充法(Conditional Mean Completer)。在该方法中,缺失属性值的补齐同样是靠该属性在其他对象中的取值求平均得到,但不同的是用于求平均的值并不是从信息表所有对象中取,而是从与该对象具有相同决策属性值的对象中取得。这两种数据的补齐方法,其基本的出发点都是一样的,以最大概率可能的取值来补充缺失的属性值,只是在具体方法上有一点不同。与其他方法相比,它是用现存数据的多数信息来推测缺失值。

(4) 热卡填充 (Hot Deck Imputation, 或就近补齐)

对于一个包含空值的对象,热卡填充法在完整数据中找到一个与它最相似的对象,然后用这个相似对象的值来进行填充。不同的问题可能会选用不同的标准来对相似进行判定。该方法概念上很简单,且利用了数据间的关系来进行空值估计。这个方法的缺点在于难以定义相似标准,主观因素较多。

(5) K 最近距离邻法 (K-means Clustering)

先根据欧式距离或相关分析来确定距离具有缺失数据样本最近的 K 个样本,将这 K 个值加权平均来估计该样本的缺失数据。

(6) 使用所有可能的值填充 (Assigning All Possible Values of the Attribute)

这种方法是用空缺属性值的所有可能的属性取值来填充,能够得到较好的补齐效果。

但是,当数据量很大或者遗漏的属性值较多时,其计算的代价很大,可能的测试方案很多。另有一种方法,填补遗漏属性值的原则是一样的,不同的只是从决策相同的对象中尝试所有的属性值的可能情况,而不是根据信息表中所有对象进行尝试,这样能够在一定程度上减小原方法的代价。

(7) 组合完整化方法 (Combinatorial Completer)

这种方法是用空缺属性值的所有可能的属性取值来试,并从最终属性的约简结果中选择最好的一个作为填补的属性值。这是以约简为目的的数据补齐方法,能够得到好的约简结果;但是,当数据量很大或者遗漏的属性值较多时,其计算的代价很大。另一种称为条件组合完整化方法 (Conditional Combinatorial Complete),与填补遗漏属性值的原则是一样的,不同的只是从决策相同的对象中尝试所有的属性值的可能情况,而不是根据信息表中所有对象进行尝试。条件组合完整化方法能够在一定程度上减小组合完整化方法的代价。在信息表包含不完整数据较多的情况下,可能的测试方案将巨增。

(8) 回归 (Regression)

基于完整的数据集,建立回归方程 (模型)。对于包含空值的对象,将已知属性值代入方程来估计未知属性值,以此估计值来进行填充。当变量不是线性相关或预测变量高度相关时会导致有偏差的估计。

(9) 期望值最大化方法 (Expectation Maximization, EM)

EM 算法是一种在不完全数据情况下计算极大似然估计或者后验分布的迭代算法。在每一迭代循环过程中交替执行两个步骤:E 步 (Expectation Step, 期望步),在给定完全数据和前一次迭代所得到的参数估计的情况下计算完全数据对应的对数似然函数的条件期望;M 步 (Maximization Step, 极大化步),用极大化对数似然函数以确定参数的值,并用于下步的迭代。算法在 E 步和 M 步之间不断迭代直至收敛,即两次迭代之间的参数变化小于一个预先给定的阈值时结束。该方法可能会陷入局部极值,收敛速度也不是很快,并且计算很复杂。

(10) 多重填补 (Multiple Imputation, MI)

多重填补方法分为三个步骤。

为每个空值产生一套可能的填补值,这些值反映了无响应模型的不确定性;每个值都被用来填补数据集中的缺失值,产生若干个完整数据集。

每个填补数据集都用针对完整数据集的统计方法进行统计分析。

对来自各个填补数据集的结果进行综合,产生最终的统计推断,这一推断考虑到了由于数据填补而产生的不确定性。该方法将空缺值视为随机样本,这样计算出来的统计推断可能受到空缺值的不确定性的影响,该方法的计算也很复杂。

(11) C4.5 方法

通过寻找属性间的关系来对遗失值填充。它寻找之间具有最大相关性的两个属性,其中没有遗失值的一个称为代理属性,另一个称为原始属性,用代理属性决定原始属性中的遗失值。这种基于规则归纳的方法只能处理基数较小的名词型属性。

就几种基于统计的方法而言,删除元组法和平均值法差于 Hot Deck、EM 和 MI;回归是比较好的一种方法,但仍比不上 Hot Deck 和 EM;EM 缺少 MI 包含的不确定成分。值得注意的是,这些方法直接处理的是模型参数的估计而不是空缺值预测本身。它们合适处理无监督学习的问题,而对有监督学习来说,情况就不尽相同了。例如,可以删除包含空值

的对象用完整的数据集来进行训练,但预测时你却不能忽略包含空值的对象。另外,C4.5 和使用所有可能的值填充方法也有较好的补齐效果,人工填写和特殊值填充则是一般不推荐使用的。

补齐处理只是将未知值补以我们的主观估计值,不一定完全符合客观事实,在对不完备信息进行补齐处理的同时,或多或少地改变了原始的信息系统。而且,对空值不正确的填充往往将新的噪声引入数据中,使挖掘任务产生错误的结果。因此,在许多情况下,还是希望在保持原始信息不发生变化的前提下对信息系统进行处理,这就是第三种方法。

3. 不处理

直接在包含空值的数据上进行数据挖掘。这类方法包括贝叶斯网络和人工神经网络等。

贝叶斯网络是用来表示变量间连接概率的图形模式,它提供了一种自然的表示因果信息的方法,用来发现数据间的潜在关系。在这个网络中,用节点表示变量,有向边表示变量间的依赖关系。贝叶斯网络仅适合于对领域知识具有一定了解的情况,至少对变量间的依赖关系较清楚的情况。否则直接从数据中学习贝叶斯网的结构不但复杂性较高(随着变量的增加,指数级增加),网络维护代价昂贵,而且它的估计参数较多,为系统带来了高方差,影响了它的预测精度。当在任何一个对象中的缺失值数量很大时,存在指数爆炸的危险。

人工神经网络可以有效地对付空值,但人工神经网络在这方面的研究还有待进一步深入展开,这里就不再介绍了。

20.2 SPSS 缺失值分析

SPSS 中缺失值的处理方法有多种,如直接删除缺失值、替换缺失值等操作,SPSS 界面菜单分析 (Analyze) 下的缺失值分析 (Missing Value Analysis) 子菜单是专门用于进行缺失值分析的模块。

20.2.1 缺失值分析过程的参数设置

SPSS 中有专门的模块来进行缺失值的操作,选择菜单“Analyze 缺失值分析 (Missing Value Analysis)”,则弹出如图 20-1 所示的对话框,此对话框即为由于进行缺失值分析操作的参数设置,主界面的各个部分具体功能如下。

1. 变量选择设置

图 20-1 中左边为待分析变量列表,其他各变量框功能如下所述。

- 定量变量 (Quantitative Variables): 用于选入进行缺失值分析的定量变量。
- 分类变量 (Categorical Variables): 用于选入进行缺失值分析的分类变量,其下的最大类别 (Maximum Categories) 选项栏用于指定分类变量允许的最多分类数量,默认为 25。
- 个案标签 (Case Labels): 用于选入对结果进行标识的变量。
- 使用所有变量 (Use All Variables): 单击此按钮会自动将左侧变量列表框中所有变量选入特定的分析列表框中,其中数值型变量选入定量变量 (Quantitative Variables) 选项框中;字符型变量选入分类变量 (Categorical Variables) 选项框中。

图 20-1 中右边的估算 (Estimation) 选项栏用于选择计算均值、相关矩阵、协方差矩阵等统计量, 其下的各个选项框功能如下。

- 成列 (Listwise): 当分析中的任意一个因变量或者分组变量中有缺失值, 则此观测记录不会被用来进行分析。
- 成对 (Pairwise): 只有计算时用到的变量有缺失值时, 则此观测记录不会被用来进行分析。
- EM: 使用 EM 方法来进行缺失值的替代。
- 回归 (Regression): 使用多元线性回归算法估计缺失值。

2. 模式 (Patterns) 设置

单击图 20-1 中的“模式 (Patterns)”按钮, 则弹出如图 20-2 所示的对话框, 此对话框用来进行关于变量的缺失值样式表格式的设置等操作。



图 20-1 “缺失值分析的 (Missing Value Analysis) 参数设置”对话框



图 20-2 “模式 (Patterns) 设置”对话框

显示 (Display) 选项栏: 此栏用于选择缺失值样式表的格式, 各选项功能如下所述:

- 按缺失值模式分组的个案表 (Tabulated Cases): 为分析变量输出缺失值的样式表, 缺失值用“X”来表示。其下的省略个案数不足多少的变量 (Omit Patterns with less than): 用于指定省略比例, 出现频数小于此比例的缺失模式将不被显示; 按照缺失值模式将变量排序 (Sort Variables by Missing Value Pattern): 用于标识按照缺失值模式排序。
- 按缺失值模式排序的具有缺失值的个案 (Cases with Missing Values): 用于为每个含有缺失值的记录给出缺失值样式表。其下的按缺失值模式将变量排序 (Sort Variables by Missing Value Pattern): 用于标识按照缺失值模式排序。

- 按选定变量指定顺序排序的所有个案 (All cases, optionally sorted by selected variable): 用于列出所有记录的缺失值情况。

变量 (Variables) 选项栏: 选中输出 (Display) 选项后, 则激活此选项栏。

- 缺失模式 (Missing Patterns for) 选项: 显示选入的所有分析变量。
- 附加信息 (Additional Information for): 用于输出所列变量的观测值列表。
- 排序依据 (Sort by): 用于指定输出观测列表的排序变量。其下的排列顺序 (Sort Order) 选项栏用于指定其排序方式为 Ascending (升序) 和 Descending (降序)。

3. 描述 (Descriptives) 设置

单击图 20-1 中的“描述 (Descriptives)”按钮, 则弹出如图 20-3 所示的对话框, 此对话框用于设置一些描述性统计量, 各个选项栏的功能具体如下。

单变量统计量 (Univariate Statistics): 用于为每个变量输出非缺失数据的个数、均值、标准差等基本统计量信息。

指示符变量统计量 (Indicator Variable Statistics): 用于为分析变量生成指示变量, 标识相应数值是否缺失, 其下的选项具体功能如下。

- 不匹配百分比 (Percent Mismatch): 表示对于每对变量, 输出其中一个变量缺失、另一个未缺失的记录所占的比例, 其下的排序依据 (Sort by) 选项框用于表示按照缺失值样式进行排序。
- 使用指示符变量形成的组进行 t 检验 (t Tests with Groups Formed by Indicator Variables): 根据指示变量标识将记录分为两组, 并对每个数值变量进行 t 检验。
- 生成分类变量和指示符变量的交叉表 (Cross Tabulations of Categorical and Indicator Variables): 为分类变量和指示变量生成交叉表。
- 省略缺失值占个案数的比例小于 (Omit Variables Missing Less Than): 指定忽略比例, 缺失值频数小于此比例的变量将不再显示。

4. 变量 (Variables) 设置

选中图 20-1 右边的 EM、回归 (Regression) 选项框之后则会激活其下的 Variables、EM、回归 (Regression) 三个按钮, 单击“变量 (Variables)”按钮, 则会弹出如图 20-4 所示的对话框, 各组成部分具体功能如下。

变量 (Variables) 选项栏: 用于选择指定变量的方式。

- 使用所有定量变量 (Use all Quantitative Variables): 表示使用所有连续变量。
- 选择变量 (Select Variables): 由用户指定分析变量, 选中以后则会激活其下的选项栏。

定量变量 (Quantitative Variables) 选项栏: 用于显示所有可用于缺失值估计的连续变量。

预测变量 (D) (Predicted Variables): 用于选入需要估计缺失值的变量。

预测变量 (R) (Predictor Variables): 用于选入用来估计其他变量缺失值的变量。

两者都包含 (Both): 单击此按钮, 则可以把定量变量 (Quantitative Variables) 列表里选中的变量, 选入预测变量 (D) (Predicted Variables) 列表框和预测变量 (R) (Predictor Variables) 列表框。

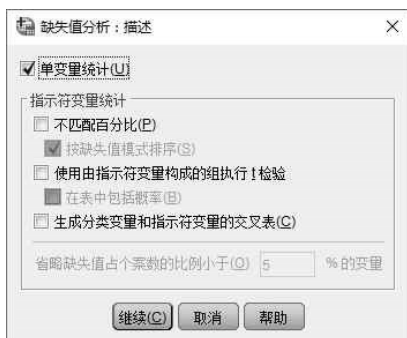


图 20-3 “描述 (Descriptives) 选项设置”对话框



图 20-4 “变量 (Variables) 选项设置”对话框

5. EM 设置

单击图 20-1 右边的“EM”按钮，则弹出如图 20-5 所示的对话框，此对话框用来设置 EM 算法的相关参数。

分布 (Distribution) 选项：用于选择总体分布的形式。

- 正态分布 (Normal)
- 混合正态 (Mixed Normal)，混合比例 (Mixture Proportion) 选项框用于设置混合比例；标准差比率 (Standard Deviation Ratio) 选项框用于设置标准差比率。
- 学生 t：学生 t 分布，自由度 (Degrees of Freedom) 选项框用于设置自由度。

最大迭代次数 (Maximum Iterations)：设置最大迭代次数，默认为 25。

保存完成的数据 (Save Completed Data)：用于保存将缺失值用 EM 算法替换后的数据。创建新数据集 (Create a New Dataset) 用于新创建一个数据集来存储数据；写入新数据文件 (Write A New Data File) 用于新建一个数据文件来存储数据，单击“文件 (File)”按钮选择路径。

6. 回归 (Regression) 设置

单击图 20-1 右边的“回归 (Regression)”按钮，则弹出如图 20-6 所示的对话框，此对话框用来设置回归 (Regression) 算法的相关参数。

估算调整 (Estimation Adjustment)：用于指定随机项的分布。

- 残差 (Residuals)：随机误差项的分布和由方程导出的残差项相同。
- 普通变量 (Normal Variables)
- 学生 t 分布变量：该栏下的自由度 (Degree of Freedom) 选项框用于设置自由度。
- 无 (None)：不添加随机项。

最大预测变量数 (Maximum Number of Predictors)：指定回归方程自变量的最大个数。

保存完整数据 (Save completed data)：保存数据文件，同 EM 选项框中设置方法。



图 20-5 “EM 参数设置”对话框

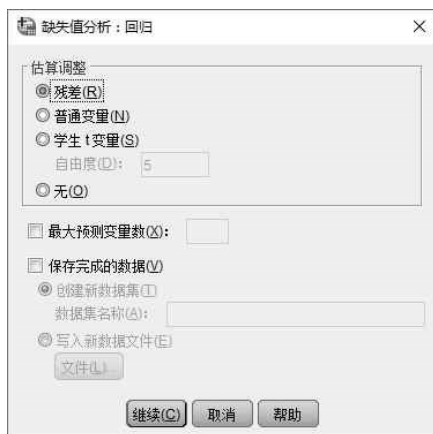


图 20-6 “回归 (Regression) 参数设置”对话框

20.2.2 实例分析



结果文件

——附带光盘“PROGRAM\CH20\实例 20-1”文件夹



动画演示

——附带光盘“AVI\实例 20-1.avi”文件

本节所用数据集为 SPSS 自带的数据集 telco_missing.sav，此数据集中有缺失值，在进行深度的数据分析之前需要对其进行分析。数据格式如图 20-7 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|----|----------------|----|----|------|----|------------------|----|----|----|----|----|
| 1 | MonthsWith... | 数字 | 4 | 0 | | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 2 | Age | 数字 | 4 | 0 | | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 3 | MaritalStatus | 数字 | 4 | 0 | | {0, Unmarrie... | 无 | 7 | 右 | 名义 | 输入 |
| 4 | YearsAtAdd... | 数字 | 4 | 0 | | 无 | 无 | 7 | 右 | 标度 | 输入 |
| 5 | Income | 数字 | 8 | 2 | | 无 | 无 | 10 | 右 | 标度 | 输入 |
| 6 | Educational... | 数字 | 4 | 0 | | {1, Did not c... | 无 | 6 | 右 | 有序 | 输入 |
| 7 | YearsWithE... | 数字 | 4 | 0 | | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 8 | Retirement... | 数字 | 8 | 2 | | {00, No} | 无 | 10 | 右 | 名义 | 输入 |
| 9 | Gender | 数字 | 4 | 0 | | {0, Male} | 无 | 6 | 右 | 名义 | 输入 |
| 10 | PeopleInHo... | 数字 | 4 | 0 | | 无 | 无 | 6 | 右 | 标度 | 输入 |

图 20-7 数据集 telco_missing.sav 格式

数据集共包括 10 个变量，即 MonthsWithService、Age、MaritalStatus、YearsAtAddress、Income、EducationalLevel、YearsWithEmployer、RetirementStatus、Gender，以及 PeopleInHousehold。数据集含有大量的缺失值。

1. 基本分析参数设置

依次选择菜单“分析 (Analyze) 缺失值分析 (Missing Value Analysis)”，则打开“缺失值分析”对话框，如图 20-8 所示。然后依次把变量 MonthsWithService、Age、YearsAtAddress、Income、YearsWithEmployer、PeopleInHousehold 选入“定量变量 (Quantitative Variables)”变量框中，把 MaritalStatus、EducationalLevel、RetirementStatus、

Gender 选入“分类变量 (Categorical Variables)”变量框中。

设置好上述一些变量以后, 就可以进行分析了, 为了更深度地进行分析, 则继续进行设置。单击“描述 (Descriptives)”按钮, 则弹出如图 20-9 所示的对话框, 并选中“使用由指示变量形成的分组进行的 t 检验 (t Tests with Groups Formed by Indicator Variables)”选项、“为分类变量和指示变量生成交叉表 (Cross Tabulations of Categorical and Indicator Variables)”选项, 然后单击“继续”按钮返回主界面。



图 20-8 “缺失值分析”对话框

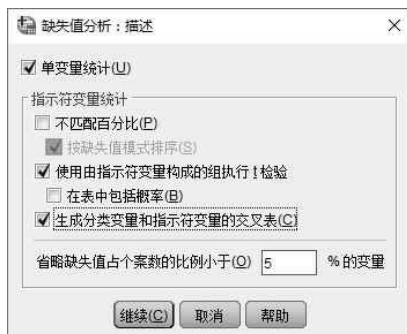


图 20-9 “描述设置”对话框

然后单击“缺失值分析”对话框界面中的“确定”按钮, 以进行缺失值分析。

2. 基本分析输出结果

此例中的输出结果包括基本统计量、单独方差 t 检验表, 以及分组变量缺失值统计信息等信息。如图 20-10 所示为分析统计量结果输出。包括观测值个数, 均值, 标准差, 以及缺失值的个数、百分比等信息。例如, 变量 Income 中含有的缺失值占全部数据的 17.9%, 变量 Age 中含有缺失值占全部数据的 2.5%。

| 单变量统计 | | | | | | | |
|-------------------|-----|---------|----------|----------|------|------------------|----|
| | 个案数 | 平均值 | 标准差 | 缺失 计数 | 百分比 | 极值数 ^a | |
| MonthsWithService | 968 | 35.56 | 21.268 | 32 | 3.2 | 0 | 0 |
| Age | 975 | 41.75 | 12.573 | 25 | 2.5 | 0 | 0 |
| YearsAtAddress | 850 | 11.47 | 9.965 | 150 | 15.0 | 0 | 9 |
| Income | 821 | 71.1462 | 83.14424 | 179 | 17.9 | 0 | 71 |
| YearsWithEmployer | 904 | 11.00 | 10.113 | 96 | 9.6 | 0 | 15 |
| PeopleInHousehold | 966 | 2.32 | 1.431 | 34 | 3.4 | 0 | 33 |
| MaritalStatus | 885 | | | 115 | 11.5 | | |
| EducationalLevel | 965 | | | 35 | 3.5 | | |
| RetirementStatus | 916 | | | 84 | 8.4 | | |
| Gender | 958 | | | 42 | 4.2 | | |

a. 超出范围 ($Q1 - 1.5 * IQR$, $Q3 + 1.5 * IQR$) 的个案数。

图 20-10 分析统计量结果输出

然后输出的是单独方差 t 检验表, 如图 20-11 所示, 这里需要注意的是当缺失值个数的比例超过 5% 时其指示变量才被创建。

从图 20-11 可以看出变量 Income 缺失值所对应的变量 Age 的均值为 49.73, 变量 Income 没有缺失值的观测所对应的变量 Age 的均值为 40.01,

| | | Months with service | age | Years at address | income | Years with employer | People in household |
|-------------------------------|----------|------------------------|-------|---------------------|---------|------------------------|------------------------|
| Years at address | t | .4 | .3 | . | 3.5 | 1.4 | 1.0 |
| | 自由度 | 202.2 | 192.5 | . | 313.6 | 191.1 | 199.5 |
| | 存在数 | 819 | 832 | 850 | 693 | 766 | 824 |
| | 缺失数 | 149 | 143 | 0 | 128 | 138 | 142 |
| | 平均值 (存在) | 35.68 | 41.79 | 11.47 | 74.0779 | 11.20 | 2.34 |
| | 平值 (缺失) | 34.91 | 41.49 | . | 55.2734 | 9.86 | 2.21 |
| income | t | -5.0 | -8.3 | -3.9 | . | -5.9 | 3.6 |
| | 自由度 | 249.5 | 222.8 | 191.1 | . | 203.3 | 315.2 |
| | 存在数 | 793 | 801 | 693 | 821 | 741 | 792 |
| | 缺失数 | 175 | 174 | 157 | 0 | 163 | 174 |
| | 平均值 (存在) | 33.93 | 40.01 | 10.67 | 71.1462 | 9.91 | 2.39 |
| | 平值 (缺失) | 42.97 | 49.73 | 14.97 | . | 15.93 | 2.02 |
| Years with employer | t | -1.0 | -.4 | -.7 | .5 | . | -.3 |
| | 自由度 | 110.5 | 110.2 | 97.6 | 114.9 | . | 110.9 |
| | 存在数 | 877 | 881 | 766 | 741 | 904 | 874 |
| | 缺失数 | 91 | 94 | 84 | 80 | 0 | 92 |
| | 平均值 (存在) | 35.34 | 41.69 | 11.37 | 71.4953 | 11.00 | 2.31 |
| | 平值 (缺失) | 37.70 | 42.27 | 12.32 | 67.9125 | . | 2.37 |
| Marital status | t | .0 | 1.8 | 1.2 | -.8 | .9 | -2.2 |
| | 自由度 | 148.1 | 149.5 | 138.8 | 121.2 | 128.3 | 134.2 |
| | 存在数 | 856 | 862 | 748 | 728 | 805 | 857 |
| | 缺失数 | 112 | 113 | 102 | 93 | 99 | 109 |
| | 平均值 (存在) | 35.56 | 42.00 | 11.61 | 70.3887 | 11.10 | 2.28 |
| | 平值 (缺失) | 35.57 | 39.85 | 10.43 | 77.0753 | 10.17 | 2.61 |
| Retirement status | t | -.6 | -.4 | -.4 | .3 | . | .2 |
| | 自由度 | 95.4 | 94.4 | 84.0 | 93.2 | . | 99.0 |
| | 存在数 | 888 | 893 | 777 | 751 | 904 | 885 |
| | 缺失数 | 80 | 82 | 73 | 70 | 0 | 81 |
| | 平均值 (存在) | 35.44 | 41.70 | 11.42 | 71.3356 | 11.00 | 2.32 |
| | 平值 (缺失) | 36.89 | 42.29 | 11.96 | 69.1143 | . | 2.30 |
| 对于每个定量变量, 由指示符变量构成组对 (存在与缺失)。 | | | | | | | |
| a. 不会显示缺失百分比低于 5% 的指示符变量 | | | | | | | |

图 20-11 单独方差 t 检验表

分组变量缺失值统计信息表, 如图 20-12 所示, 此表给出了和单独方差 t 检验表相似的信息。

| | | | 总 计 | Unmarried | Married | 缺 失 |
|---------------------|----|----------|------|-----------|---------|-------|
| | | | | | | 系统缺失值 |
| Years at address | 存在 | 计数 | 850 | 390 | 358 | 102 |
| | | 百分比 | 85.0 | 85.5 | 83.4 | 88.7 |
| | 缺失 | 系统缺失值百分比 | 15.0 | 14.5 | 16.6 | 11.3 |
| income | 存在 | 计数 | 821 | 380 | 348 | 93 |
| | | 百分比 | 82.1 | 83.3 | 81.1 | 80.9 |
| | 缺失 | 系统缺失值百分比 | 17.9 | 16.7 | 18.9 | 19.1 |
| Years with employer | 存在 | 计数 | 904 | 418 | 387 | 99 |
| | | 百分比 | 90.4 | 91.7 | 90.2 | 86.1 |
| | 缺失 | 系统缺失值百分比 | 9.6 | 8.3 | 9.8 | 13.9 |
| Retirement status | 存在 | 计数 | 916 | 423 | 392 | 101 |
| | | 百分比 | 91.6 | 92.8 | 91.4 | 87.8 |
| | 缺失 | 系统缺失值百分比 | 8.4 | 7.2 | 8.6 | 12.2 |

不会显示缺失百分比少于 5%的指示符变量

图 20-12 分组变量缺失值统计信息表

3. 深度分析设置

下面进一步进行分析，单击图 20-8 中的“模式 (Patterns)”按钮，则弹出如图 20-13 所示对话框，此对话框用于设置变量模式。如图 20-13 所示选择“按照缺失值模式分组的表格个案 (Tabulated Cases)”选项栏，以及其下面的选项“个案表 (按缺失值模式分组) (Grouped by Missing Values Patterns)”，以及把变量 Income、EducationalLevel、RetirementStatus、Gender 选入“附加信息 (Additional information for)”变量框中。因为变量 Income 的缺失值很多，所以有必要给出更加详细的分析结果，最后单击“继续 (Continue)”按钮返回主界面。

4. 深度分析输出结果

单击主界面“缺失值分析 (Missing Value Analysis)”对话框中的“确定”按钮，则系统输出分析的结果，如图 20-14 为模式输出结果。



图 20-13 “模式 (Patterns) 设置”对话框

| | | 案 例 数 | | | | | | | | | | |
|----------------------------------|------------------------------|---------|-----|-----|---------|---------|---------|---------|---------|---------|-----|---------|
| | | 475 | 109 | 16 | 87 | 13 | 60 | 16 | 17 | 18 | 16 | 37 |
| 缺失模式 ^a | age | | | | | | | | | | | |
| | People in household | | | | | X | | | | | | |
| | Months with service | | | | | | | | X | | | |
| | Educational level | | | | | | | X | | | | |
| | gender | | | | | | | | | X | | |
| | Retirement status | | | | | | | | | | | X |
| | Years with employer | | | | | | | | | | | X |
| | Marital status | | | | | | X | | | | X | |
| | Years at address | | | X | X | | | | | | | |
| | income | | X | X | | | | | | | X | |
| 完成条件 ^b | | 475 | 584 | 687 | 562 | 488 | 535 | 491 | 492 | 493 | 660 | 520 |
| income ^c | | 76.5853 | . | . | 54.4368 | 56.0000 | 77.2167 | 47.8125 | 76.2353 | 54.1111 | . | 59.4595 |
| Educational level ^d | Did not complete high school | 99 | 27 | 5 | 21 | 4 | 1 | 0 | 2 | 3 | 0 | 9 |
| | High school degree | 157 | 35 | 9 | 27 | 3 | 2 | 0 | 7 | 7 | 0 | 14 |
| | Some college | 87 | 19 | 0 | 9 | 2 | 27 | 0 | 3 | 4 | 7 | 5 |
| | College degree | 101 | 17 | 1 | 24 | 3 | 24 | 0 | 4 | 4 | 8 | 8 |
| | Post-undergraduate degree | 31 | 11 | 1 | 6 | 1 | 6 | 0 | 1 | 0 | 1 | 1 |
| Retirement status ^d | No | 463 | 95 | 12 | 85 | 13 | 59 | 16 | 17 | 17 | 14 | 0 |
| | Yes | 12 | 14 | 4 | 2 | 0 | 1 | 0 | 0 | 1 | 2 | 0 |
| gender ^d | Male | 201 | 47 | 12 | 66 | 4 | 35 | 6 | 7 | 0 | 6 | 15 |
| | Female | 274 | 62 | 4 | 21 | 9 | 25 | 10 | 10 | 0 | 10 | 22 |
| 将不会显示个案百分比低于 1% (10 个或更少) 的模式 | | | | | | | | | | | | |
| a. 变量按缺失模式排列； | | | | | | | | | | | | |
| b. 不使用模式 (以 X 标记) 中的缺失变量时的完整个案数； | | | | | | | | | | | | |
| c. 每个唯一模式的平均值； | | | | | | | | | | | | |
| d. 每个唯一模式的频率分布 | | | | | | | | | | | | |

图 20-14 模式表输出结果

从图 20-14 中可以看出实例中变量缺失值发生超过 1% 有三个缺失值模式，最后一行可以得到变量 YearsWithEmployer 和 RetirementStatus 常常比其他成对变量更加的同时发生缺失。所以，变量 RetirementStatus 和 YearsWithEmployer 的记录信息基本一致就显的不奇怪

了。从变量 Income 一行中可以看出变量 Income 的均值比变量 MaritalStatus 缺失时的均值高出 6%，同时也比变量 MonthsWithService 缺失时的均值高。

最后进行利特尔 MCAR 检验分析，选中图 20-8 中的 EM 选项栏，然后单击“确定”则给出输出结果，如图 20-15、图 20-16、图 20-17 所示，给出了 EM 算法估计后的均值、协方差和相关性的表格。

| Months with service | age | Years at address | income | Years with employer | People in household |
|---------------------|-------|------------------|---------|---------------------|---------------------|
| 36.12 | 41.91 | 11.58 | 77.3941 | 11.22 | 2.29 |

a. 利特尔的 MCAR 检验：卡方 = 179.836，自由度 = 107，显著性 = .000

图 20-15 EM 检验结果

| | Months with service | age | Years at address | income | Years with employer | People in household |
|---------------------|---------------------|---------|------------------|------------|---------------------|---------------------|
| Months with service | 460.893 | | | | | |
| age | 135.326 | 161.261 | | | | |
| Years at address | 111.341 | 85.440 | 105.372 | | | |
| income | 547.182 | 451.109 | 300.533 | 7664.75710 | | |
| Years with employer | 113.359 | 86.871 | 48.051 | 525.81159 | 103.326 | |
| People in household | -1.107 | -4.538 | -3.098 | -14.60886 | -1.916 | 2.006 |

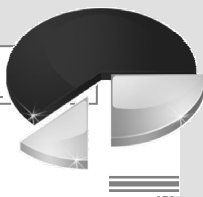
a. 利特尔的 MCAR 检验：卡方=179.836，自由度=107，显著性=.000

图 20-16 EM 协方差

| | Months with service | age | Years at address | income | Years with employer | People in household |
|---------------------|---------------------|-------|------------------|--------|---------------------|---------------------|
| Months with service | 1 | | | | | |
| age | .496 | 1 | | | | |
| Years at address | .505 | .655 | 1 | | | |
| income | .291 | .406 | .334 | 1 | | |
| Years with employer | .519 | .673 | .461 | .591 | 1 | |
| People in household | -.036 | -.252 | -.213 | -.118 | -.133 | 1 |

a. 利特尔的 MCAR 检验：卡方=179.836，自由度=107，显著性=.000

图 20-17 EM 相关性



第 21 章 决策树模型

SPSS Classification Trees 附加模块，能够在 SPSS 环境下直接创建分类决策树，帮助你快速并准确地识别群体，发现群体之间的关系并预测未来事件。可应用分类决策树于分段、分层、预测、数据降维、变量筛选、类别合并，以及连续变量离散化。

本章将具体介绍 SPSS 中决策树的详细应用。



本讲内容

- 决策树模型概述
- SPSS 中参数设置
- 实例分析

21.1 决策树模型概述

决策树 (Decision Tree) 一般都是自上而下来生成的。每个决策或事件 (自然状态) 都可能引出两个或多个事件，导致不同的结果，把这种决策分支画成图形很像一棵树的枝干，故称决策树。

决策树主要有如下几个优点。

- 可以生成可以理解的规则。
- 计算量相对来说不是很大。
- 可以处理连续和种类字段。
- 决策树可以清晰地显示哪些字段比较重要。

但是也有如下缺点。

- 对连续性的字段比较难预测。
- 对有时间顺序的数据，需要很多预处理的工作。
- 当类别太多时，错误可能就会增加得比较快。
- 一般的算法分类的时候，只是根据一个字段来分类。

决策树的一般分类过程如下：在树根处从样本数据观测值中选择一个目标变量 (Target)，例如，本文中的黑色样本和白色样本标识，以及用来分离 (Split) 样本观测的输入

变量 (Input); 然后在树枝上逐一选择输入变量, 对目标变量进行分离, 并且计算衡量同一分类中样本同质性和不同分类中样本异质性的指标, 选择能够最大化类间异质性和类中同质性的分类, 所有输入变量的分组结束之后, 选择其中某一个能够最大化类间异质性和类中同质性的输入变量, 将其作为树枝上的分类规则, 把样本分离到不同的内部节点; 然后在每一个内部节点上重复上一步骤, 直到所有的内部节点只包含同一类样本为止, 将最后的节点作为树叶。

决策树根据特定的算法, 如 χ^2 检验 (Chi-square Test)、最大熵减少量 (Entropy Reduction)、基尼系数减少量等 (Gini Reduction), 自动从样本中收集信息, 从树根开始, 不断选取新的属性来区分样本, 直到所有内部节点中的样本都被区分到某个类别中。

SPSS 中提供分类树算法有四种算法, 使能够尝试不同类型的树生成方法, 并找到最佳拟合数据的模型, 具体算法如下。

- CHAID——快速、多分枝的统计树算法, 使能够迅速有效地探索数据, 可根据所希望的分类结果建立分段及资料概括说明。
- Exhaustive CHAID——改进的 CHAID 算法, 会检查预测因子的每种可能分割。
- Classification and Regression Trees (CRT)——一个完全的二叉树算法, 能将数据分割为精确、类似同质的子集合。
- QUEST——可以无偏差地选择变量, 迅速有效地建立二叉树的算法。

表 21-1 给出了不同类型的决策树算法及适应性等信息。

表 21-1 决策树算法及适应性等信息

| 算 法 | 特 性 | 优 点 |
|------------------|--|---|
| CHAID | 非参数。 定向转换自变量不影响结果。 处理特别数点较好。 可用于连续变量和类别变量的任何组合。 可以调整样本。 可以提出变量的相互作用。 适用二元资料的判定树。 分割规格为 GainRatio。 修剪规格为错误预估率 | CHAID 会防止数据被过度套用并让判定树停止继续分割, 依据的衡量标准是计算节点中类别的 P 值大小, 以此决定判定树是否继续分割, 所以不需要作树剪枝 |
| CRT | CRT 利用 entropy 或 GiniMetric 来选取最佳的 split。 CHAID 利用 chi square test 来决定哪一个 predictor 与所预测值越相关。 适用于非二元资料的判定树。 分割规格为卡方检定。 修剪规格为不用修剪 | CRT 算法会自动检验模型, 找出最佳的一般模型。 CRT 算法在处理遗漏资料方面是相当拿手的。 可运用于复杂数据。 可以较好处理 MISSINGVALUES。 不需事先选好变数 |
| Exhaustive CHAID | 它为修正 CHAID 而来, 主要是让所有可能切割的数据, 能更彻底的完成料的分群。 适用于非二元资料的判定树。 合并选项方面上, 若是有序的数据形态, 只能前后有序的合并而不能跳脱顺序来合并 | 利用多层的统计判定树方法作精确的数据内涵检视。 会根据用户指定项目的数量自动的划分连续的变数 |

续表

| 算 法 | 特 性 | 优 点 |
|-------|--|---|
| QUEST | <p>为了更快更有精确, Quest 分不同阶段的执行变量选择和分隔点选择。</p> <p>适用二元资料的判定树。</p> <p>QUEST 只接受名义上的目标变量</p> | <p>快速有效地建立一个正确的二元树模型。</p> <p>可以不偏的做变量选择</p> |

SPSS 中的决策树模型具有广泛的应用, 有以下几个方面。

1. 数据库营销

选择一个响应变量对客户细分 (回复/未回复邮件的; 高、中和低利润客户; 营销目标囊括那些延长服务期的客户)。

基于其他属性概括客户群, 如人口统计学或客户活动。

针对某具体客户群进行新的个性化营销, 以减少成本, 提高投资回报率 (ROI)。

2. 市场研究

进行客户、雇员, 或招募满意度的调查。

选择一个变量作为满意度的度量 (例如, 取值 “ 1 ~ 5 ” 等级变量)。

根据响应者对其他问题的回答描述满意度水平。

改善影响满意度的因素, 如工作环境或产品质量。

3. 信用风险分析

确定风险组 (高、中或低)。

基于客户信息描述风险群体特征, 如账户活动。

基于风险群为贷款申请人提供合理的贷款额度。

4. 项目目标

选择变量用以表示是否发生预期结果 (例如, 一福利项目的成功实施)。

根据申请人信息, 揭示决定成功的因素。

开展新的项目, 以满足更多人的需要。

5. 公共部门营销

选择一个响应变量划分你的客户群 (例如, 高校潜在申请人中实际申请者与未申请者)。

基于其他属性概括客户群, 如人口统计学或客户活动。

针对某具体客户群进行新的个性化营销, 以减少成本, 提高投资回报率 (ROI)。

21.1.1 CHAID 算法

CHAID 算法模型是一种分类方法, 它的理论构想主要来源于决策树模型, 决策树模型是以树型分类研究根据因变量在自变量上的分布来划分人群。决策树模型的算法很多, 有

CRT、C4.5 和 CHAID 等, CHAID 算法模型是现在市场研究和社会调查研究应用的比较广的方法。

CHAID 算法模型由 AID 和 thaid 这两种算法模型演变而来。AID 算法每一次只能把原来众多的变量水平划分为两类, 而往往研究中需要把一个变量的不同水平划分为多于两类, 所以, AID 算法对进一步研究形成局限, 而 Messenger 和 Mandell 发展出来的 thaid 算法没有解决结果受样本量影响的问题。CHAID 方法是 Kass 在 1975 年提出的。它的基本分析思路是 χ^2 自动交叉检验, 用 χ^2 检验免除了结果受样本大小发生变化的问题, 而且还能得到分类的显著性检验。这种方法大量应用在信件回复率的研究上, 根据信件回复情况来找出回复率较高的人群和较低的人群, 也应用在其他满意度研究或是某一类型人的特征研究中, 如宾馆的满意度研究, 赌博人员特征研究, 它的应用范围非常广, 现在 Answer Tree 和 SPSS 中都有这个算法的模块。

21.1.2 Exhaustive CHAID 算法

Exhaustive CHAID 算法是 CHAID 的改进算法, 于 1991 年由 Biggs、De Ville 等人提出 CHAID 在分箱过程中, 如果发现无须再合并就停止合并, 但 Exhaustive CHAID 将继续合并属性变量组 (依据统计量观测值大小), 最终形成两个超组, 在组的合并上较 CHAID 算法更“彻底”, 利于分支变量的选择。

21.1.3 CRT 算法

分类与回归树 CRT 最早由 Breiman 等人于 1984 提出, Ripley 在 1996 年进行了修改。变量分为预测变量 (Predict Variable) 和应变量 (因变量 (Dependent Variable)), 该模型使用二叉树将预测空间递归地划分为若干子集, 而树中的叶节点对应着划分的不同区域, 划分是由与每个内部节点相关的分支规则 (Splitting Rules) 来确定的, 通过从树根到叶节点移动, 一个预测样本被赋予一个唯一的叶节点, 应变量在该节点上的条件分布也即被确定。CRT 算法包含三部分内容: 分枝变量及拆分点的选择、树的修剪和模型树的评估。

1. 分枝变量及拆分点的选择

分类树理想的结果是使得树中每一个叶节点要么是纯节点 (节点内部样本的应变量属于同一个类), 要么很小 (节点内部所含样本个数小于事先给定的 n 值)。在从众多的预测变量中选择这个最佳分组变量时, CRT 算法采用基尼系数来进行评判。基尼系数越小, 表明该节点越纯, 则该预测变量就是当前属性的最优分割点。对基尼系数的介绍可参考有关文献。在对样本集进行分割时, 分割规则采用二叉表示形式, 算法从根节点开始分割, 递归地对每个节点重复进行。

2. 树的修剪 (Pruning)

由于数据中有噪声和孤立点, 许多分枝反映的是训练数据中的异常。CRT 采取的是后剪枝 (Postpruning) 方法, 剪去不可靠的分枝, 以提高树正确的分类能力。CRT 采用 CRT 系统的成本—复杂度最小 (Minimal Cost-complexity Pruning) 原则进行删减。

3. 评估树模型

CRT 法采用测试样本评估 (Test Sample Estimates) 交叉验证评估 (Cross-validation Estimates) 或 V-折交叉验证 (V-fold Cross-validation), 使得最终的模型树分类误判率低且树模型简单, 对于最终模型树大小的选择要结合资料的专业背景及统计结果来选择。

21.1.4 QUEST 算法

QUEST 所运用的算法相当的技术化, 但分类树模块亦提供分裂 (分层) 选取方法选项提供用户较简易的计算方法。QUEST 技巧着重于快速且不偏, 但预测变数较多, 则指令周期远较 CRT 来的有弹性许多 (Loh 与 Shih 于 1997 年所提出的实验报告中指出, QUEST 花费 1CPU 秒, 而 CRT 却必须要花费 30.5 CPU 秒)。当部分预测变量拥有较少的等级时, QUEST 的偏误缺适度侦测 (不偏性), 对于分裂 (分层) 的找寻也提供相当好的结果。且其他变量有较多的等级数时, 最后的结果证实, QUEST 并不会因处理速度较快而牺牲预测精确性。

21.2 决策树的参数设置

在 SPSS 中可以利用决策树 (Classification Tree) 过程来实现决策树模型, 选择菜单“分析 (Analyze) 分类 (Classify) 树 (Tree)”, 则系统执行决策树分析过程, 弹出如图 21-1 所示对话框, 此对话框主要是告诉用户在进行分类树分析之前一定要正确设置分析变量的度量方式。

单击如图 21-1 所示对话框中的“确定 (OK)”按钮后, 则弹出如图 21-2 所示对话框, 为分类树的主界面。

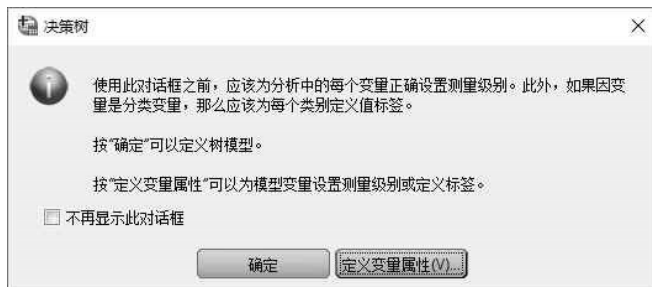


图 21-1 “说明”对话框

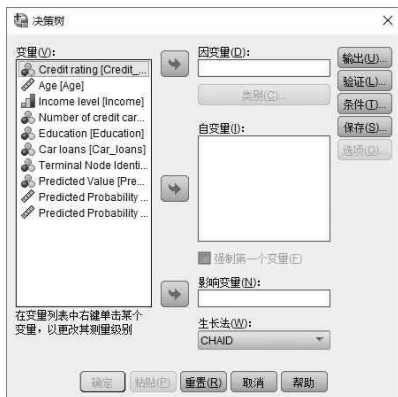


图 21-2 分类树主界面

21.2.1 变量设置

进行分析设置之前要进行变量的选择设置, 图 21-2 的左边为待分析的变量列表, 其他各变量含义如下所述。

1. 因变量（因变量（Dependent Variable））选项栏

用于从变量列表中选入一个因变量。

2. 自变量（In 因变量（Dependent Variable）s）选项栏

用于从变量列表中选入多个自变量。

3. 强制使用第一个变量（Force First Variable）选项栏

表示直接将自变量选项栏中的第一个变量，作为决策树生长的开始节点的分枝变量。

4. 影响变量（Influence Variable）选项栏

用于从变量列表框中选入一个影响自变量。该变量反映单个观测对决策树生长的影响程度，取值越大影响越大，必须为数值型的变量，且不能设置因变量为影响变量，如果指定了 QUEST 算法，将忽略此变量。

5. 生长法（Growing Method）下拉菜单

用于指定决策树的生长算法，各算法如图 21-3 所示。

- CHAID：卡方自动交互检测法。
- 穷举 CHAID（Exhaustive CHAID）：改进的 CHAID 算法。
- CRT：分类回归树算法。
- QUEST：快速、无偏、有效的统计数算法。

21.2.2 类别（Categories）设置

选入变量到“因变量（Dependent Variable）”选项栏后，则激活其下的类别“（Categories）”按钮，单击此按钮则弹出如图 21-4 所示对话框。用于设置因变量目标取值定义对话框。



图 21-3 决策树的生长算法



图 21-4 “类别（Categories）设置”对话框

变量（Variable）：Credit rating，用于显示当前的因变量名称。

在分析中使用（Use in Analysis）选项栏。

- 类别 (Category): 给出当前因变量的值标签。
- 目标 (Target): 给出一列复选框, 用于选择目标取值。
- 排除 (Exclude) 选项栏: 用于选入不参与分析的因变量取值。

21.2.3 输出 (Output) 设置

单击图 21-2 中的“输出 (Output)”按钮, 则弹出如图 21-5 所示的对话框, 用于设置输出的一些参数, 各部分组成如下所述。

1. 树 (Tree) 界面

用于输出图形决策树。

- 输出 (Display) 选项栏: 设置图形决策树的输出格式。方向 (Orientation) 表示显示方向; 节点内容 (Node Contents) 表示节点的内容; 标度 (Scale) 表示显示范围; 自变量统计信息 (Independent Variable Statistics) 表示对于 CHAID 和 Exhaustive CHAID 算法, 要求在节点中显示连续变量的 F 统计量、显著性水平, 以及自由度, 分类变量的卡方统计量、显著性水平及其自由度。对于 CRT 算法, 显示每一步的改进值。对于 QUEST 算法, 显示连续变量和有序变量的 F_{tjil}、显著性水平及其自由度等统计信息。节点定义 (Node Definitions) 表示节点定义, 显示父节点分支时所用的自变量在其每个子节点的取值。
- 使用表格式的树 (Tree in Table Format): 以表格形式输出决策树, 包括每个节点的节点统计信息等内容。

2. 统计量 (Statistics) 界面

单击图 21-5 中的“统计量 (Statistics)”标签, 则弹出如图 21-6 所示对话框, 用于设置各种统计量。

模型 (Model) 选项栏。

- 摘要 (Summary): 摘要信息, 包括模型的方法等信息。
- 风险 (Risk): 风险估计及其标准误, 用来衡量决策树的预测精度。
- 分类表 (Classification Table): 分类表。对于分类因变量给出其每个取值水平上的判断正确数和错误数。对于连续因变量, 不作任何输出。
- 成本、先验概率、得分和利润值 (Cost, prior probability, score and profit values): 对于分类因变量, 输出错判损失函数, 先验概率, 得分和分析所使用的得益函数。对于连续因变量, 不作任何输出。
- 节点性能 (Node Performance) 选项栏: 用于设置关于节点的统计信息。
- 摘要 (Summary): 摘要表格输出。
- 按目标类别 (By Target Category): 对于定义了目标取值的分类因变量, 此表包括得益比例、相应比例、以节点或者百分比分组后的增量 (Lift) 值, 对每个目标取值输出一个表格, 对于连续因变量和没有定义目标的分类因变量不作输出。
- 自变量 (Independent Variables) 选项栏: 用于设置自变量的选项。

- 对模型的重要性 (Importance to Model): 对于 CRT 方法, 把模型中的自变量按其重要性进行排序, 对其他算法无效。
- 替代变量 (按拆分) (Surrogates by Split): 对于 CRT 和 QUEST 算法, 如果模型有可替代的解决方案, 就列出所有可能的方案, 对 CHAID 算法无效。

行 (Row) 下拉列表, 用于指定节点信息表的显示方式, 可以选择终端节点 (Terminal Nodes)、百分位数 (Percentiles) 和两者都是 (Both)。如果选择两者都是, 则为因变量的每个目标取值的输出两个表格。百分表按指定顺序依次显示指定百分位处的累计值。

- 排序顺序 (Sort Order): 用于指定百分位表的显示顺序。
- 百分位数增量 (Percentile Increment): 在此指定百分位的递增间隔。
- 显示累积统计信息 (Display Cumulative Statistics): 表示在每个最终节点表里增加一列显示累计结果。



图 21-5 “输出 (Output) 设置”对话框



图 21-6 “统计量 (Statistics) 界面设置”对话框

3. 规则 (Rules) 界面

单击图 21-5 中的“规则 (Rules)”标签, 则弹出如图 21-7 所示的对话框, 用于设置输出的一些参数, 各部分组成如下。

生成分类规则 (Generate Classification Rules) 选项栏: 表示输出分类决策规则。

语法 (Syntax) 选项栏: 用于设置关于决策规则的语句格式。

- SPSS Statistics: 输出 SPSS 命令语句。
- SQL: 输出标准的 SQL 语句。
- 简单文本 (Simple Text): 简单文本输出。

类型 (Type) 选项栏: 用于设置关于语法 (Syntax) 和 SQL 格式的决策规则的类型。

- 为个案指定值 (Assign Values to Cases): 生成一条符合节点规则的赋值语句。
- 选择个案 (Select Cases): 生成一条符合节点规则的选择语句。
- 将替代变量包含在 SPSS Statistics 和 SQL 规则中 (Include Surrogates in SPSS and SQL Rules)。

节点 (Nodes) 选项栏。

- 所有终端节点 (All Terminal Nodes): 对每个最终节点输出规则。
- 最佳终端节点 (Best Terminal Nodes): 其下的节点数 (Number of Nodes) 用于指定 n 的值。
- 达到指定个案百分比的最佳终端节点数 (Best Terminal Nodes up to a Specified Percentage of Cases): 其下的百分比 (Percentage) 输入框用于指定 n 的数值。
- 索引值满足或超过分界值的终端节点 (Terminal Nodes Whose Index Value Meets or Exceeds a Cut Off Value): 其下的最小索引值 (Minimum) 输入框用于指定 n 的数值。
- 所有节点 (All Nodes): 对所有节点都输出决策规则。

将规则导出至文件 (Export Rules to a File): 设置把决策规则输出至指定的文件之中。单击“浏览 (Browse)”按钮指定文件路径。



图 21-7 “输出 (Output) 规则设置”对话框

21.2.4 验证 (Validation) 设置

单击图 21-2 中的“验证 (Validation)”按钮，则弹出如图 21-8 所示对话框，此对话框是验证设置子界面，用来判断模型的稳定性和通用性。

无 (None) 选项：不进行验证。

交叉验证 (Cross Validation) 选项：交叉验证，其下的样本群数 (Number of Sample Folds) 表示子样本个数的整数，不超过 25 个。

分割样本验证 (Split-sample Validation) 选项：表示样本分离验证。此方法将样本分为两个子集，即训练样本和验证样本，利用训练样本拟合决策树模型，利用验证样本检验模型。

- 使用随机分配 (Use Random Assignment): 随机划分，其下的训练样本 (Training Sample) (%) 表示用于指定训练集占总体的比例，验证集占总体的比例显示在 Test Sample 之后。

- 使用变量 (Use Variable): 通过指定变量来划分数据集, 变量 (Variables) 选项显示可用的变量, 样本拆分依据 (Split Sample By) 表示用于选入划分数据集的变量, 值为 1 的个案将分配给训练样本。所有其他个案将用在检验样本中。显示以下项的结果 (Display Results For) 选项栏: 此栏用于设置哪些样本输出分析结果。
- 训练和检验样本 (Training and Test Sample): 表示对训练集和验证集都输出相关的结果。
- 仅检验样本 (Test Sample Only): 表示只对验证集输出有关结果。



图 21-8 “验证 (Validation) 设置”对话框

21.2.5 保存 (Save) 设置

单击如图 21-2 中的“保存 (Save)”按钮, 则弹出如图 21-9 所示的对话框, 各部分选项功能如下所述。

1. 已保存变量 (Saved Variables) 选项栏

用于设置保存哪些变量。

- 终端节点数 (Terminal Node Number): 表示节点序号, 此变量保存每个观测所属最终节点的序号。
- 预测值 (Predicted Value): 此变量保存由模型预测的因变量值。
- 预测概率 (Predicted Probabilities)
- 样本分配 (训练/检验) (Sample Assignment (Training/Testing)): 样本类型, 此变量记录单个观测是用于训练函数用于验证。

2. 将树模型以 XML 导出 (Export Tree Model as XML) 选项

设置把模型格式输出到指定 XML 文件的选项。

- 训练样本 (Training Sample): 设置对训练样本的输出。
- 检验样本 (Test Sample): 设置对验证样本的输出。



图 21-9 “保存 (Save) 设置”对话框

21.2.6 条件 (Criteria) 设置

单击图 21-2 中的“条件 (Criteria)”按钮，则弹出如图 21-10 所示对话框，此对话框用于设置各种算法的参数，不同的算法对应不同的界面，首先介绍算法 CHAID 和 Exhaustive CHAID 的界面。



图 21-10 “条件 (Criteria) 设置”对话框

最大树深度 (Maximum Tree Depth) 选项栏：用于设置决策树在根节点以下的最大深度。

- 自动 (Automatic): 自动设置算法的深度，对于算法 CHAID 和 Exhaustive CHAID，最大深度为 3；对于算法 CRT 和 QUEST，最大深度为 5。
- 定制 (Custom): 用户自定义，其下的值 (Value) 用来设置深度。

最小个案数 (Minimum Number of Cases) 选项栏：用于设置每个节点需要的最少观测个数。

- 父节点 (Parent Node)：指定父节点需要的最少观测数量，默认为 100。
- 子节点 (Child Node)：指定子节点需要的最少观测数量，默认为 50。

21.2.7 CHAID 算法设置

单击图 21-10 中的 CHAID 标签，则弹出如图 21-11 所示对话框，此对话框用于设置算法 CHAID 的各种参数。



图 21-11 “CHAID 算法设置”对话框

以下项的显著性水平 (Significance Level For) 选项栏：其下有两个选项。

- 拆分节点 (Splitting Nodes)：指定分割节点的显著性水平临界值，系统默认为 0.05。
- 合并类别 (Merging Categories)：指定合并节点的显著性水平临界值，系统默认为 0.05。

模型估算 (Model Estimation) 选项栏：对于分类因变量，可以设置如下的参数。

- 最大迭代次数 (Maximum Number of Iterations)：用来指定最大的迭代次数，默认数值为 100。
- 期望的单元格频率中的最小更改 (Minimum Change in Expected Cell Frequencies)：指定单元格频数的最小改变量。

卡方统计 (Chi-Square Statistics) 选项栏：用于设置卡方统计量。

- 皮尔逊 (Pearson)：皮尔逊，系统默认选项。
- 似然比 (Likelihood Ratio)：似然比卡方。

使用 Bonferroni 方法调整显著性值 (Adjust Significance Values Using Bonferroni Method)：用 Bonferroni 方法调整合并或者分割节点时的显著性水平，默认选项。

允许重新拆分节点中合并后的类别 (Allow Resplitting of Merged Categories within a Node)：表示合并的节点进行重新分割以生成更好的决策树。

21.2.8 CRT 算法设置

在图 21-2 中的“增长方法 (Growing Methods)”下拉菜单中，选择 CRT 选项，然后单击“条件 (Criteria)”按钮，则弹出如图 21-12 所示对话框，用于设置 CRT 算法的参数。

杂质测量 (Impurity Measure) 选项栏：用于设置节点内部的 Impurity 度量。

- 基尼选项 (Gini)：寻求使子节点内部的因变量一致性达到最高的分支方法。
- 两分法 (Twoing) 选项：把因变量的取值水平分为两个子集，寻求使这两个子集分得最开的方案。
- 顺序两分法 (Ordered Twoing) 选项：与上面的两分法 (Twoing) 类似，但是要求只有因变量相邻的取值才可以合并为一类。

改进中的最小更改 (Minimum Change in Improvement) 选项：指定分割一个节点所需要的最小不纯度减少值，默认为 0.0001。



图 21-12 “CRT 算法设置”对话框

21.2.9 QUEST 算法设置

在图 21-2 中的“增长方法 (Growing Methods)”下拉菜单中，选择 QUEST 选项，然后单击“条件 (Criteria)”按钮，则弹出如图 21-13 所示对话框，用于设置 QUEST 算法的参数。

界面中只有一个设置选项框，即拆分节点的显著性水平 (Significance Level for Splitting Nodes) 选项栏，用于指定一个大于 0 小于 1 的数值，系统默认为 0.05。

21.2.10 修剪 (Pruning) 设置

单击图 21-12 中的“修剪 (Pruning)”标签，则弹出如图 21-14 所示对话框，用于设置算法的剪枝参数。

1. 修剪树以避免过度拟合 (Prune tree to avoid over fitting) 选项栏

此栏表示决策树长满以后，对其进行修剪一面生长过度。

2. 风险中的最大差分 (标准误差) (Maximum Difference in Risk) 选项栏

用于指定决策树被剪枝前后所允许的风险值的最大差额，以标准差的方式表示，系统默认为 1。增大此值，将生成更小的决策树，设为 0 时，将输出风险最小的决策树。



图 21-13 “QUEST 算法设置”对话框

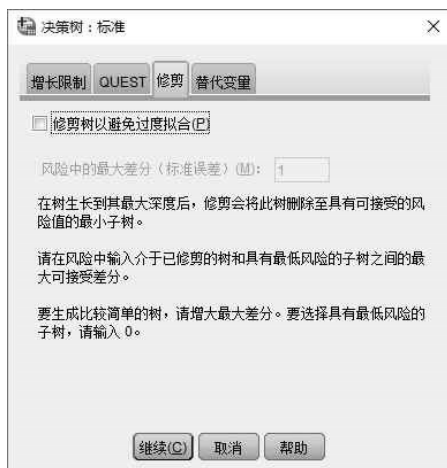


图 21-14 “修剪 (Pruning) 设置”对话框

21.2.11 替代变量 (Surrogates) 设置

单击图 21-14 中的“替代变量 (Surrogates)”标签，即弹出如图 21-15 所示对话框，用于设置关于算法的被选方案的参数。

只有一个“最大替代变量数 (Maximum Number of Surrogates)”选项栏，用于指定模型中允许使用的备选自变量的最大个数。

- 自动 (Automatic): 系统默认项。
- 定制 (Custom): 用户自定义，其后的“值 (Value)”选项框用于填入自变量的最大个数。

21.2.12 选项 (Options) 设置

单击图 21-2 中的“选项 (Options)”按钮，则弹出如图 21-16 所示的对话框，选项 (Options) 共有四个标签，首先是缺失值 (Missing Values) 标签，如图 21-16 所示。



图 21-15 “替代变量 (Surrogates) 设置”对话框

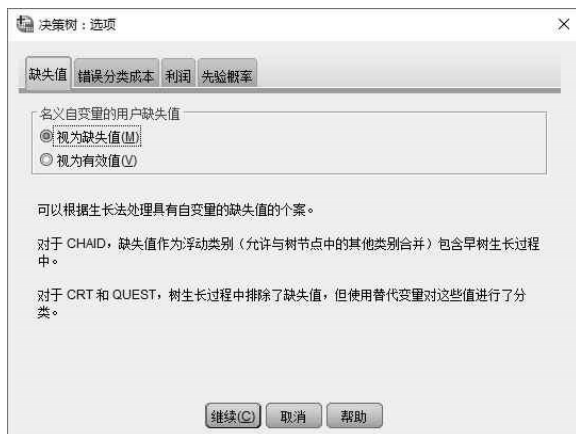


图 21-16 “缺失值 (Missing Values) 设置”对话框

其界面中的名义自变量的用户缺失值 (User-Missing Values of Nominal Independent Variables) 选项栏, 用于设置名义自变量的用户定义缺失值如何处理。

- 视为缺失值 (Treat as Missing Values): 当做系统缺失值处理。
- 视为有效值 (Treat as Valid Values): 当做正常的有效值处理。

这里要注意的是对于不同的算法, 对缺失值的处理方法也不同。

21.2.13 错误分类成本设置

单击图 21-16 中的“错误分类成本 (Missclassification Costs)”标签, 则弹出如图 21-17 所示的对话框, 用于设置关于错判惩罚函数的参数。



图 21-17 “错误分类成本 (Missclassification Costs) 设置”对话框

在各类别之间相等 (Equal Across Categories) 选项栏: 用于表示各种错判分类的惩罚都是一样的。

定制 (Custom) 选项栏: 用户自定义错判惩罚函数。当分类因变量设置了两个值标签时, 此项才可用。

填充矩阵 (Fill Matrix) 选项栏: 设置如何使惩罚矩阵成为对称的形式。

- 复制下三角形 (Duplicate Lower Triangle): 把矩阵的下三角复制到上三角使之对称。
- 复制上三角形 (Duplicate Upper Triangle): 把矩阵的上三角复制到下三角使之对称。
- 使用平均单元格值 (Use Average Cell Values): 计算任意两个对称单元格的算术平均值并取代它们, 使矩阵对称。

21.2.14 利润 (Profits) 设置

单击图 21-16 中的“利润 (Profits)”标签, 则弹出如图 21-18 所示的对话框, 用于设置预测分类正确时的收益函数的参数。

无 (None): 不使用收益函数。

定制 (Custom) 选项栏: 表示由用户自定义收益函数。

只有当分类因变量至少设置了两个值标签时, 此选项栏才可用。收入 (Revenue) 表示输入对当前行的值标签预测正确时的收入值; 费用 (Expense) 表示对当前行的值标签预测正确时的消耗值; 利润 (Profit) 表示收益值。



图 21-18 “利润 (Profits) 设置”对话框

21.2.15 先验概率 (Prior Probabilities) 设置

单击图 21-16 中的“先验概率 (Prior Probabilities)”标签，则弹出如图 21-19 所示的对话框，用于设置先验概率的有关参数，各组成部分如下。

从训练样本 (经验先验) 中获取 (Obtain from Training Sample (empirical prior)) 选项栏：用于指定先验概率。从训练样本集中获得先验概率，当样本能很好的代表总体时选中此项。此项为系统默认选项。

在各类别之间相等 (Equal Across Categories) 选项栏：用于指定等先验概率。当因变量各取值水平所占的比例都很相近时，选中此项。

定制 (Custom) 选项栏：用户自定义先验概率，其下的二维表中 Value 用于输入当前行标签所对应的先验概率。

使用错误分类成本调整先验 (Adjust Priors using Misclassification Costs) 选项栏：如果定义了错判惩罚函数，可以选中此项，表示用错判矩阵对先验概率进行调整。



图 21-19 “先验概率 (Prior Probabilities) 设置”对话框

21.2.16 实例分析

本实例主要利用分类树来研究信用风险，所用数据集为 SPSS 自带的数据库文件 tree_credit.sav。



结果文件——附带光盘“PROGRAM\CH21\实例 21-1”文件夹



动画演示——附带光盘“AVI\实例 21-1.avi”文件

假设银行有一个记载客户取得贷款交易信息的数据库，包括客户偿还或拖欠贷款的记录。使用分类树技术，银行方面可以分析及时还贷和有拖欠行为的客户特征，并能建立模型预测后续的贷款申请者拖欠银行贷款的可能性。

数据文件 tree_credit.sav 的数据格式，如图 21-20 所示。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|----|----------------|----|----|------|---------------------|------------------|------|----|----|----|----|
| 1 | Credit_rating | 数字 | 8 | 2 | Credit rating | {00, Bad}... | 9.00 | 13 | 右 | 名义 | 输入 |
| 2 | Age | 数字 | 8 | 2 | Age | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 3 | Income | 数字 | 8 | 2 | Income level | {1.00, Low}... | 无 | 8 | 右 | 有序 | 输入 |
| 4 | Credit_cards | 数字 | 8 | 2 | Number of credit... | {1.00, Less ...} | 无 | 12 | 右 | 名义 | 输入 |
| 5 | Education | 数字 | 8 | 2 | Education | {1.00, High ...} | 无 | 9 | 右 | 名义 | 输入 |
| 6 | Car_loans | 数字 | 8 | 2 | Car loans | {1.00, None ...} | 无 | 9 | 右 | 名义 | 输入 |
| 7 | ModelID | 数字 | 8 | 0 | Terminal Node I... | 无 | 无 | 8 | 右 | 名义 | 输入 |
| 8 | PredictedVa... | 数字 | 8 | 2 | Predicted Value | 无 | 无 | 8 | 右 | 名义 | 输入 |
| 9 | PredictedPr... | 数字 | 8 | 2 | Predicted Prob... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 10 | PredictedPr... | 数字 | 8 | 2 | Predicted Prob... | 无 | 无 | 8 | 右 | 标度 | 输入 |

图 21-20 数据集 tree_credit.sav 的数据格式

21.2.17 模型建立

本实例中使用 CHAID 算法。在计算的每一步中，CHAID 选择与因变量交互作用最强的自变量（预测因子）。如果某些自变量与因变量没有很强的显著性差别，这些自变量的分类将被合并。

1. 建立 CHAID 树模型

选择菜单“分类 (Analyze) 分类 决策树 (Tree)”，然后弹出如图 21-21 所示对话框，选中变量 Credit rating 到“因变量 (Dependent Variable)”选项栏中，作为因变量。选中变量 Age、Income level、Number of credit cards、Education、Car loans 到“自变量”选项栏，作为自变量。

2. 选择分类目标

单击图 21-21 中的“类别 (Categories)”按钮，则弹出如图 21-22 所示对话框，在这里就可以指定感兴趣的目标分类。目标分类自身不影响树模型，但是如果选择了目标分类，部分输出和选项就可使用。在分类为 Bad 的目标复选框中打钩。具有不良信用等级（拖欠贷款）的客户将按感兴趣的目标分类来对待。然后单击“继续 (Continue)”按钮返回主界面。



图 21-21 “树 (Tree) 设置”对话框



图 21-22 “类别 (Categories) 设置”对话框

3. 定义生成标准

单击图 21-21 中的“条件 (Criteria)”按钮，然后弹出如图 21-23 所示对话框，在最小个案数 (Minimum Number of Cases) 组中，父节点处输入 400，子节点处输入 200。然后单击“继续 (Continue)”按钮返回主界面。

4. 选择附加输出

在分类树对话框中单击“输出 (Output)”按钮。出现一个多页对话框，如图 21-24 所示，在这里可以选择各种附加输出类型。

选择树 (Tree) 页的“表格树 (Tree in Table format)”选项。然后单击图 (Plots) 页，弹出如图 21-25 所示对话框，选择“增益 (Gain)”和“索引 (Index)”选项。然后单击“继续 (Continue)”按钮返回主界面。

5. 保存预测值

保存包含模型预测信息的变量。例如，保存每个个案预测的信用等级，然后与实际信用等级进行比较。



图 21-23 “条件 (Criteria) 设置” 对话框



图 21-24 “输出 (Output) 设置” 对话框

在“树设置”对话框中单击“保存 (Save)”按钮。弹出如图 21-26 所示的对话框，选择“终端节点数 (Terminal Node Number)”、“预测值 (Predicted Value)”和“预测概率 (Predicted probabilities)”选项。然后单击“继续 (Continue)”按钮返回主界面。



图 21-25 “图 (Plots) 设置” 对话框



图 21-26 “保存 (Save) 设置” 对话框

21.2.18 模型评估

设置好上述的参数以后，则单击主界面的“确定 (OK)”按钮进行分类树分析。本案例中，模型结果包括以下几个方面。

- 提供有关模型信息的表格。
- 树形图。
- 提供模型性能指示的图表。
- 将模型的预测变量添加到当前工作的数据文件中。

1. 模型汇总表

模型汇总表提供有关建立模型的一些信息，如图 21-27 所示。指定（Specifications）部分提供产生树模型设置的信息，包括生成方法为 CHAID，因变量为信用等级，自变量为年龄，收入，信用卡数，教育及汽车贷款。有效性验证为没有，最大树深度是 3，父节点中最小个案是 400，子节点中最小个案是 200。

结果（Results）部分显示在最终模型中选入的自变量为年龄，收入和信用卡数。总节点数为 10，端点数为 6，树的深度（根节点下的树叶数）为 3。

有 5 个自变量被选入，但是最终模型只选中 3 个。变量教育和汽车贷款对模型没有显著的贡献，所以，它们自动地从最终模型中排出。

| 模型摘要 | | |
|------|------------|---|
| 指定项 | 生长法 | CHAID |
| | 因变量 | Credit rating |
| | 自变量 | Age, Income level, Number of credit cards, Education, Car loans |
| | 验证 | 无 |
| | 最大树深度 | 3 |
| | 父节点中的最小个案数 | 400 |
| | 子节点中的最小个案数 | 200 |
| 结果 | 包括的自变量 | Income level, Number of credit cards, Age |
| | 节点数 | 10 |
| | 终端节点数 | 6 |
| | 深度 | 3 |

图 21-27 模型汇总表

2. 树形图

树形图是树模型的图解表示。树模型显示如图 21-28 所示。本树形图使用的是 CHAID 方法，收入水平是信用等级的最佳预测因子。

观察低收入那一枝（子节点 1），收入水平是与信用等级唯一有显著意义的因子。在这个类别中有 82% 银行客户（Bad）拖欠贷款。只有 18% 的客户（Good）按时还贷。由于节点 1 下面没有子节点，节点 1 就是端点。

然后观察中高收入客户群（子节点 2 和子节点 3），信用卡数是它的最佳因子。

子节点 4 为有 5 张以上信用卡的中等收入客户群，它还包括另一个预测因子：年龄。年龄在 28 岁以下的 80.8% 的客户有不良信用等级，它几乎是 28 岁以上组的不良信用等级数（43.7%）的两倍。

可以使用 Tree Editor 隐藏和显示选择的树枝，改变颜色和字体，依据选择的节点选择个案的子集，在此不再累述。

3. 树表

如它的名字一样，树表以表格的形式提供大部分实用的树形图信息。对每个节点，表

的显示如图 21-29 所示。输出的主要是因变量在每个分类中个案的数量和百分比。

从图中可以得到因变量的预测分类。在本例中, 预测分类标准是按照所在节点的个案数超过 50% 来进行信用等级分类, 因为只有两个可能的信用等级, 将个案数的百分率低于 50% 的划归 Bad 组, 高于 50% 的划归 Good 组。如节点 1, 不良对良好组的比例是 82.1% 对 17.9%, 所以节点 1 的预测分类为 Bad。节点 2, 不良对良好组的比例是 42.0% 对 58.0%, 所以节点 2 的预测分类为 Good。依此类推。

parent node 表示树中每个节点的父节点。注意节点 1 (低收入节点) 不是任何节点的父节点, 因为它是端点, 没有子节点。节点 4、节点 5 的父节点是节点 2, 节点 6、节点 7 的父节点是节点 3, 节点 8、节点 9 的父节点是节点 4。

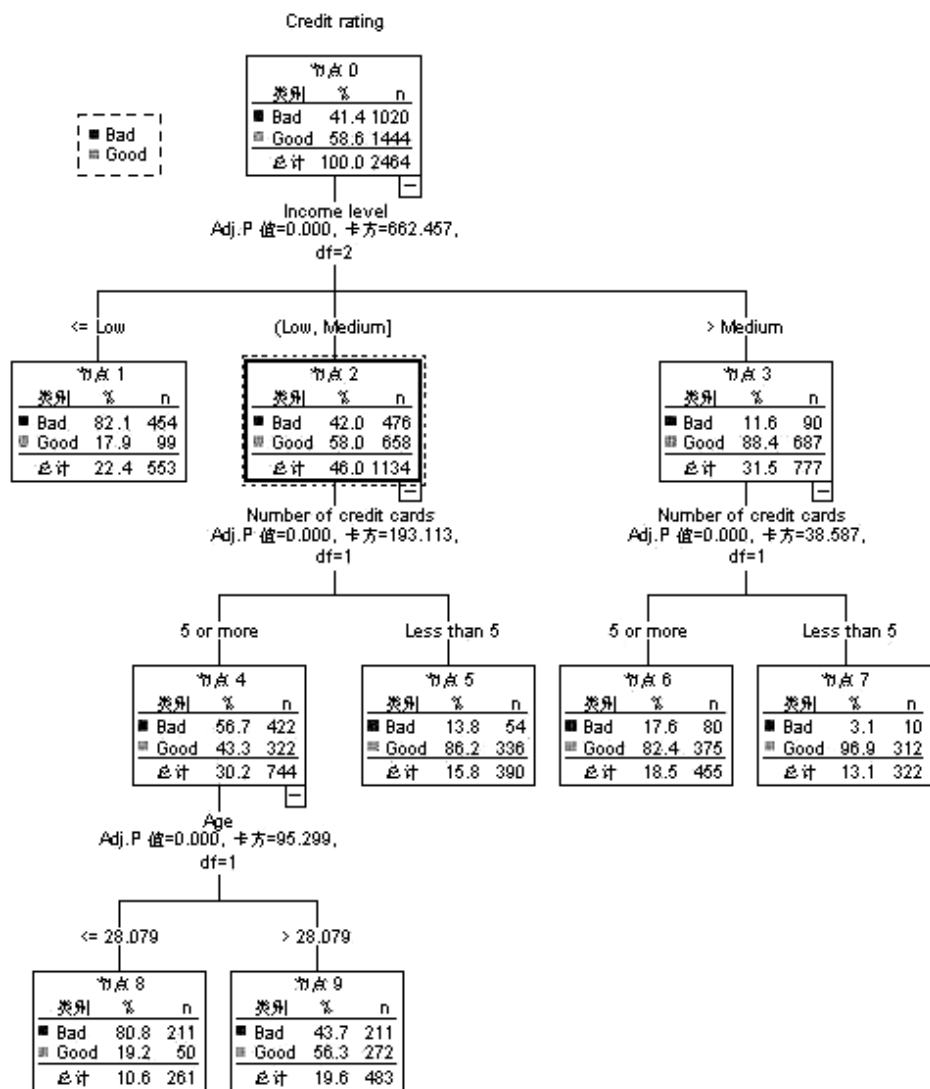


图 21-28 树形图

自变量用来划分节点。如节点 1, 节点 2, 节点 3 由收入水平划分, 节点 4, 节点 5, 节点 6, 节点 7 由信用卡数来划分, 节点 8, 节点 9 由年龄来划分。

4. 节点增益

节点表的增益提供了模型中端点的汇总信息，如图 21-30 所示。只有端点在这张表中列出。通常只对端点感兴趣，因为它们表达了模型中最好的分类预测。

| | | 节 点 | | | | | | | | | |
|--------------------|------------------|--------|-----------------|------------------|-----------------|------------------------------|------------------------------|------------------------------|------------------------------|--------------|-------------|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Bad | 个案数 | 1020 | 454 | 476 | 90 | 422 | 54 | 80 | 10 | 211 | 211 |
| | 百分比 | 41.4% | 82.1% | 42.0% | 11.6% | 56.7% | 13.8% | 17.6% | 3.1% | 80.8% | 43.7% |
| Good | 个案数 | 1444 | 99 | 658 | 687 | 322 | 336 | 375 | 312 | 50 | 272 |
| | 百分比 | 58.6% | 17.9% | 58.0% | 88.4% | 43.3% | 86.2% | 82.4% | 96.9% | 19.2% | 56.3% |
| 总计 | 个案数 | 2464 | 553 | 1134 | 777 | 744 | 390 | 455 | 322 | 261 | 483 |
| | 百分比 | 100.0% | 22.4% | 46.0% | 31.5% | 30.2% | 15.8% | 18.5% | 13.1% | 10.6% | 19.6% |
| 预测类别 | | Good | Bad | Good | Good | Bad | Good | Good | Good | Bad | Good |
| 父节点 | | | 0 | 0 | 0 | 2 | 2 | 3 | 3 | 4 | 4 |
| 主要 自变 量 | 变量 | | Income level | Income level | Income level | Number of credit cards | Number of credit cards | Number of credit cards | Number of credit cards | Age | Age |
| | 显著性 ^a | | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 |
| | 卡方 | | 662.457 | 662.457 | 662.457 | 193.113 | 193.113 | 38.587 | 38.587 | 95.299 | 95.299 |
| | 自由度 | | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 拆分值 | | <= Low | (Low, Medium] | > Medium | 5 or more | Less than 5 | 5 or more | Less than 5 | <= 28.079 | > 28.079 |
| 生长法：CHAID | | | | | | | | | | | |
| 因变量：Credit rating | | | | | | | | | | | |
| a . Bonferroni 已调整 | | | | | | | | | | | |

图 21-29 树表

| 节点的增益 | | | | | | |
|-------------------|-----|-------|-----|-------|-------|--------|
| 节点 | 节点 | | 增益 | | 响应 | 指数 |
| | 个案数 | 百分比 | 个案数 | 百分比 | | |
| 1 | 553 | 22.4% | 454 | 44.5% | 82.1% | 198.3% |
| 8 | 261 | 10.6% | 211 | 20.7% | 80.8% | 195.3% |
| 9 | 483 | 19.6% | 211 | 20.7% | 43.7% | 105.5% |
| 6 | 455 | 18.5% | 80 | 7.8% | 17.6% | 42.5% |
| 5 | 390 | 15.8% | 54 | 5.3% | 13.8% | 33.4% |
| 7 | 322 | 13.1% | 10 | 1.0% | 3.1% | 7.5% |
| 生长法：CHAID | | | | | | |
| 因变量：Credit rating | | | | | | |

图 21-30 节点增益输出

因为增益值提供有关目标分类的信息, 这张表只有在指定一个或多个目标分类时才可用。本例中, 只有一个目标分类, 所以节点表只有一个增益。

节点 N 表示每个端点的总个案数, 节点百分数是每个节点总个案数除以根节点的总个案数。如节点 1 的总个案数为 553, 节点百分数为 $553/2464=22.4\%$ 。

Gain N 是每个端点在目标分类中所标记的个案数。Gain Percent 是目标分类的个案数除以该类在总分类中的个案数, 本例中, 由于选择不良信用等级为感兴趣的分类, 所以, 增益就表示具有不良信用等级的个案的数和百分数。如节点 1 Bad 组的个案数为 454, 节点 8 Bad 组的个案数为 211, 而根节点 Bad 组的总个案数是 1020, 所以, 节点 1 的百分数是 $454/1020=44.5\%$, 节点 8 的百分数是 $211/1020=20.7\%$ 。

对分类因变量而言, Response 为目标分类中个案的百分数。本例中, 感兴趣的是 Bad 组, 所以, 显示各端点的 Bad 组的百分数, 如节点 1 为 82.1%, 节点 2 为 80.8%。

索引是目标分类的响应百分比除以总样本中该类的响应百分比。如节点 1 Bad 组的百分比为 82.1%, 根节点中 Bad 组的百分比为 41.39%, 所以, 节点 1 的索引为 $82.1/41.39=198.3\%$, 节点 8 的索引为 $80.8/41.39=195.3\%$ 。

索引值表示所在节点观测目标分类百分比与期望目标分类百分比相差多少的程度。即各节点的 Bad 组的百分比与根节点的 Bad 组的百分比的差别。

索引值大于 100%意味着各端点的 Bad 组的百分比大于根节点的 Bad 组的百分比。与此相反, 索引值小于 100%意味着各端点的 Bad 组的百分比小于根节点的 Bad 组的百分比。

5. 增益图

增益图显示模型相当好, 如图 21-31 所示。累计增加图总是从 0%开始, 到达 100%结束。好的模型, 增加图开始向 100%方向陡峭上升然后慢慢变平整。没有信息提供的模型为对角线方向的直线。

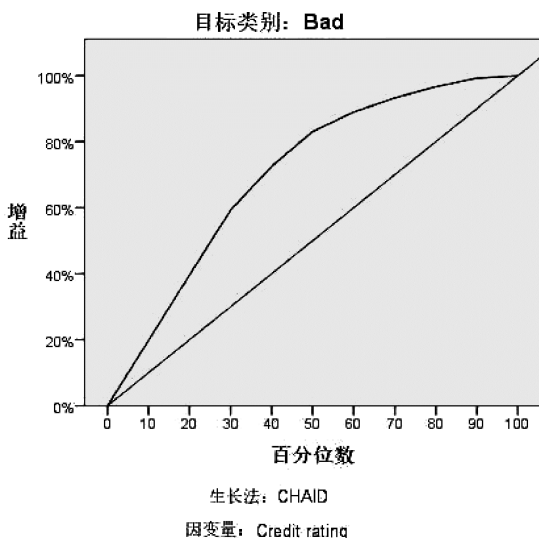


图 21-31 增益图

6. 索引图

索引图也显示该模型是一个的良好模型，如图 21-32 所示。累计索引图从 100%以上开始，渐渐地下降直到它们达 100%处。对于一个好模型来讲，索引值应该从 100%轴的上面开始，沿移动方向保持高原平坦，然后快速地向 100%下降。对没有提供信息的模型，整个图是围绕 100%盘整的直线。

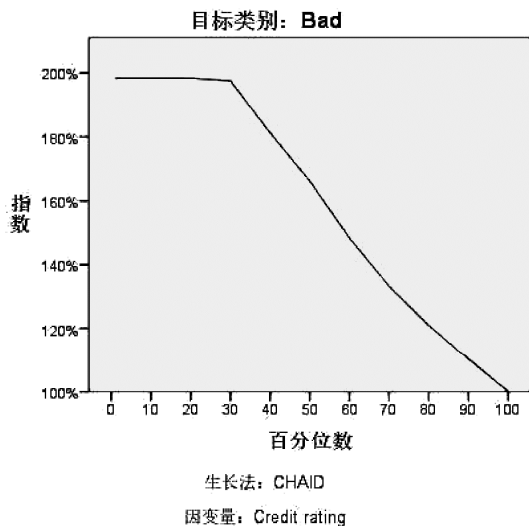


图 21-32 索引图

7. 风险评估与分类

风险和分类表可以对模型进行快速评估，如图 21-33 所示。从风险评估表中可以看出，0.205 的风险估计表示按照模型（良好和不良信用等级），预测分类错误率为 20.5%。所以，错分客户的风险约 21%。

分类表中给出的结果与风险估计一致。分类表显示客户正确地分类约 79.5%。然而，对这个模型，分类表揭示一个潜在的问题：对不良信用等级的客户，它只预测了不良信用等级的 65%，这就意味着具有不良信用等级的客户有 35%被不正确地划为信用“良好”的客户。

| 估 算 | 标 准 误 差 | | |
|-------|---------|-------|-------|
| .205 | .008 | | |
| 实 测 | 预 测 | | |
| | Bad | Good | 正确百分比 |
| Bad | 665 | 355 | 65.2% |
| Good | 149 | 1295 | 89.7% |
| 总计百分比 | 33.0% | 67.0% | 79.5% |

生长法: CHAID

因变量: Credit rating

图 21-33 风险评估与分类

8. 预测值

四个新变量已经在当前工作数据文件中建立,如图 21-34 所示。其中 NodeID 表示每个个案的端点数。PredictedValue 表示每个个案的预测值。因为因变量的编码是 0=Bad 和 1= Good, 预测值 0 意味着个案被预测为不良信用等级。PredictedProbability 表示属于每个分类的概率。因为因变量只有两个可能值,所以建立两个变量: PredictedProbability_1 表示属于不良信用等级分类个案的概率。PredictedProbability_2 表示属于良好信用等级分类个案的概率。

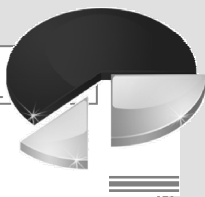
| | NodeID_1 | PredictedValue_1 | PredictedProbability_1_1 | PredictedProbability_2_1 |
|----|----------|------------------|--------------------------|--------------------------|
| 1 | 9 | 1.00 | .44 | .56 |
| 2 | 8 | .00 | .81 | .19 |
| 3 | 1 | .00 | .82 | .18 |
| 4 | 1 | .00 | .82 | .18 |
| 5 | 9 | 1.00 | .44 | .56 |
| 6 | 9 | 1.00 | .44 | .56 |
| 7 | 9 | 1.00 | .44 | .56 |
| 8 | 1 | .00 | .82 | .18 |
| 9 | 1 | .00 | .82 | .18 |
| 10 | 8 | .00 | .81 | .19 |

图 21-34 模型预测值

在包含个案端点的因变量分类中,预测概率是简单的个案比例。例如,节点 1 有 82% 的个案是不良信用等级,18% 具有良好信用等级,结果是两组预测概率分别是 0.82 和 0.18。

对分类因变量来讲,预测值划分的标准依据个案端点中具有个案最高比例。例如,第一个个案,预测值是 1 (良好信用等级),因为它的端点中大约有 56% 的个案有良好信用等级。

与此相反,对第二个个案,预测值是 0 (不良信用等级),因为它的端点中大约有 81% 的个案具有不良信用等级。



第 22 章 神经网络

人工神经网络是根据人的认识过程而开发出的一种算法。假如现在只有一些输入和相应的输出,而对如何由输入得到输出的机理并不清楚,那么,可以把输入与输出之间的未知过程看做是一个“网络”,通过不断地给这个网络输入和相应的输出来“训练”这个网络,网络根据输入和输出不断地调节自己的各节点之间的权值来满足输入和输出。这样,当训练结束后,给定一个输入,网络便会根据自己已调节好的权值计算出一个输出。这就是神经网络的简单原理。

本章利用 SPSS 神经网络工具模块,在深入浅出地介绍人工神经网络中的各种典型网络,以及训练过程的基础上,利用 SPSS 进行神经网络的设计与应用。



本讲内容

- 神经网络概述
- SPSS 神经网络模型的参数设置
- 实例分析

22.1 神经网络概述

思维学普遍认为,人类大脑的思维分为抽象(逻辑)思维、形象(直观)思维和灵感(顿悟)思维三种基本方式。

逻辑性的思维是指根据逻辑规则进行推理的过程;它先将信息化成概念,并用符号表示,然后,根据符号运算按串行模式进行逻辑推理;这一过程可以写成串行的指令,让计算机执行。然而,直观性的思维是将分布式存储的信息综合起来,结果是忽然间产生想法或解决问题的办法,这种思维方式的根本之点在于以下两点。

- 信息是通过神经元上的兴奋模式分布存储在网络上。
- 信息处理是通过神经元之间同时相互作用的动态过程来完成的。

人工神经网络就是模拟人思维的第二种方式。这是一个非线性动力学系统,其特色在于信息的分布式存储和并行协同处理。虽然单个神经元的结构极其简单,功能有限,但大量神经元构成的网络系统所能实现的行为却是极其丰富多彩的。

神经网络的研究内容相当广泛,反映了多学科交叉技术领域的特点。目前,主要的研究工作集中在以下几个方面。

生物原型研究。从生理学、心理学、解剖学、脑科学、病理学等生物科学方面研究神经细胞、神经网络、神经系统的生物原型结构及其功能机理。

建立理论模型。根据生物原型的研究,建立神经元、神经网络的理论模型。其中包括概念模型、知识模型、物理化学模型、数学模型等。

网络模型与算法研究。在理论模型研究的基础上构建具体的神经网络模型,以实现计算机模拟或准备制作硬件,包括网络学习算法的研究。这方面的工作也称为技术模型研究。

人工神经网络应用系统。在网络模型与算法研究的基础上,利用人工神经网络组成实际的应用系统,例如,完成某种信号处理或模式识别的功能、构建专家系统、制成机器人等。

首先,介绍有关人工神经网络的发展历史。

22.1.1 历史及现状

1943 年,心理学家 W.Mcculloch 和数理逻辑学家 W.Pitts 在分析、总结神经元基本特性的基础上首先提出神经元的数学模型。此模型沿用至今,并且直接影响着这一领域研究的进展。因而,他们两人可称为人工神经网络研究的先驱。

1945 年,冯·诺依曼领导的设计小组试制成功存储程序式电子计算机,标志着电子计算机时代的开始。1948 年,他在研究工作中比较了人脑结构与存储程序式计算机的根本区别,提出了以简单神经元构成的再生自动机网络结构。但是,由于指令存储式计算机技术的发展非常迅速,迫使他放弃了神经网络研究的新途径,继续投身于指令存储式计算机技术的研究,并在此领域作出了巨大贡献。虽然,冯·诺依曼的名字是与普通计算机联系在一起的,但他也是人工神经网络研究的先驱之一。

20 世纪 50 年代末,F.Rosenblatt 设计制作了“感知机”,它是一种多层的神经网络。这项工作首次把人工神经网络的研究从理论探讨付诸工程实践。当时,世界上许多实验室仿效制作感知机,分别应用于文字识别、声音识别、声纳信号识别,以及学习记忆问题的研究。然而,这次人工神经网络的研究高潮未能持续很久,许多人陆续放弃了这方面的研究工作,这是因为当时数字计算机的发展处于全盛时期,许多人误以为数字计算机可以解决人工智能、模式识别、专家系统等方面的一切问题,使感知机的工作得不到重视;其次,当时的电子技术工艺水平比较落后,主要的元件是电子管或晶体管,利用它们制作的神经网络体积庞大,价格昂贵,要在规模上与真实的神经网络相似是完全不可能的;另外,在 1968 年一本名为《感知机》的著作中指出线性感知机功能是有限的,它不能解决如异感这样的基本问题,而且多层网络还不能找到有效的计算方法,这些论点促使大批研究人员对于人工神经网络的前景失去信心。60 年代末期,人工神经网络的研究进入了低潮。

另外,在 20 世纪 60 年代初期,Widrow 提出了自适应线性元件网络,这是一种连续取值的线性加权求和阈值网络。后来,在此基础上发展了非线性多层自适应网络。当时,这些工作虽未标出神经网络的名称,但实际上就是一种人工神经网络模型。

随着人们对感知机兴趣的衰退,神经网络的研究沉寂了相当长的时间。20 世纪 80 年

代初期，模拟与数字混合的超大规模集成电路制作技术提高到新的水平，完全付诸实用化，此外，数字计算机的发展在若干应用领域遇到困难。这一背景预示，向人工神经网络寻求出路的时机已经成熟。美国的物理学家 Hopfield 于 1982 年和 1984 年在美国科学院院刊上发表了两篇关于人工神经网络研究的论文，引起了巨大的反响。人们重新认识到神经网络的威力，以及付诸应用的现实性。随即，一大批学者和研究人员围绕着 Hopfield 提出的方法展开了进一步的工作，形成了 80 年代中期以来人工神经网络的研究热潮。

22.1.2 神经网络特点

神经网络是对人脑生物神经网络的简化、抽象与模拟，是一种模仿人脑结构及功能的信息处理系统，它可呈现出人脑的许多特征，并具有人脑的一些基本功能。

1. 基本特征

(1) 结构特征

- 并行处理：神经网络是由大量简单处理元件相互连接构成的高度并行的非线性系统，具有大规模并行性处理特征。
- 分布式存储：结构上的并行性使神经网络的信息存储必然采用分布式方式，分布在网络所有的连接权中。
- 容错性：神经网络的容错性表现为两个方面：其一，网络中部分神经元损坏时不会对系统的整体性能造成影响；其二，神经网络能通过联想恢复完整的记忆，实现对不完整输入信息的正确识别。

(2) 能力特征

- 自学习能力：神经网络的自学习能力是指当外界环境发生变化时，经过一段时间的训练或感知，神经网络能通过自动调整网络结构参数，使得对于给定输入能产生期望的输出。
- 自组织能力：神经网络的自组织能力是指神经系统能在外部刺激下按一定规则调整神经网络元之间的突触连接，逐渐构建起神经网络。
- 自适应性：神经系统的自适应性是指神经系统通过改变自身的性能以适应环境变化的能力。实际上自适应性包含了自学习和自组织两层含义，它是通过自学习和自组织实现的。

2. 主要功能

人工神经网络具有人脑生物神经系统的某些智能特点。

(1) 联想记忆

神经网络具有分布存储信息和并行计算的性能，因此，它具有对外界刺激信息和输入模式进行联想记忆的能力。联想记忆又分为自联想和异联想记忆两种。

(2) 非线性映射

神经网络通过对系统输入输出样本对照进行自动学习，能够以任意精度逼近任意复杂的非线性映射。

(3) 分类与识别

由于神经网络可以很好地解决对非线性曲面的逼近,因此,对于在样本空间上区域分割曲面十分复杂的事物,神经网络具有很强的识别和分类能力。

(4) 优化计算

优化计算指在已知的约束条件下,寻找一组参数组合,使由该组合确定的目标函数达到最小值。神经网络将目标函数设计为网络的能量函数,无需对目标函数求导即可求解。神经网络的工作状态以动态系统方程描述,当系统状态趋于稳定时,神经网络方程的解作为输出优化结果。

(5) 知识处理

与人脑类似,神经网络可以从对象的输入输出信息中抽取规律而获得关于对象的知识,并将知识分布在网络的连接中予以存储。

22.1.3 神经元模型

从神经元的特性和功能可以知道,神经元是一个多输入-单输出的信息处理单元,而且,它对信息的处理是非线性的。根据神经元的特性和功能,可以把神经元抽象为一个简单的数学模型。工程上用的人工神经元模型如图 22-1 所示,它是一个多输入-单输出的非线性元件,其输入-输出关系可以描述为

$$I_i = \sum_{j=1}^n w_{ji} x_j - \theta_i$$

$$y_i = f(I_i)$$

式中, $x_j, j=1,2,\dots,n$ 是从其他细胞传递来的输入信号; θ_i 为神经元的阈值; w_{ji} 表示从细胞 j 到细胞 i 的连接权值(对于激发状态, w_{ij} 取正值;对于抑制状态, w_{ij} 取负值); n 为输入信号数目; y_j 为神经元输出; t 为时间; $f(\cdot)$ 为传递函数,有时候称为激发或激励函数,往往采用 0 和 1 二值函数或者 S 型函数,这一种函数都是连续和非线性的。

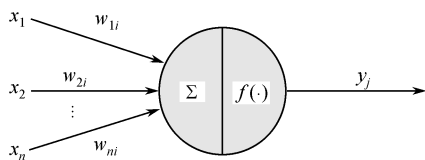


图 22-1 神经元结构模型

传递函数可以为线性函数,但是通常为像阶跃函数或者 S 状曲线那样的非线性函数。比较常用的神经元非线性函数如下。

- 阈值型函数: $f(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$
- S 状函数: $f(x) = \frac{1}{1 + e^{-\beta x}}$ 或者 $f(x) = \tanh(x)$

有时候在网络中还采用下列简单的非线性函数。

$$f(x) = \frac{x}{1 + |x|}$$

22.1.4 神经网络模型

由于 SPSS 中主要给出了径向基函数 (Radial Basis Function, RBF) 神经网络, 所以, 下面详细介绍这种神经网络模型。

RBF 神经网络是以函数逼近理论为基础而构造的一类前向网络, 这类网络的学习等价于在多维空间中寻找训练数据的最佳拟合平面, RBF 网络的每个隐层神经元传递函数都构成了拟合平面的一个基函数, 网络也由此得名。

一个典型的 RBF 神经网络由输入层、径向基层 (也称隐含层) 和输出层三层组成。它以 RBF 作为隐单元的“基”构成隐含层空间; 隐含层对输入矢量进行变换将低维的模式输入数据变换到高维空间内, 使得在低维空间内的线性不可分问题在高维空间内线性可分。通常采用高斯函数作为径向基神经元的传递函数。

如图 22-2 所示为一个具有 r 维输入的径向基函数神经元模型。图中的 dist 模块表示求取输入矢量 p 和权值矢量 w 的距离。

径向基层输入为

$$n = \sqrt{\sum (w_{li} - p_i)^2} \cdot b = \|W - P\| \cdot b$$

径向基层输出为

$$a = f(n) = e^{-n^2} = \text{radbas}(n)$$

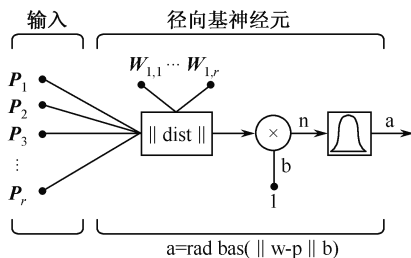


图 22-2 具有 r 维输入的径向基神经元

输出层的输出为隐含层节点输出的线性组合, 即

$$y = \text{purelin}(W' \cdot a + b')$$

式中: W' 为隐含层到输出层的权重; b' 为输出层的阈值。

22.1.5 神经网络的学习规则

神经网络的学习规则可以粗略分成三类。

第一类学习规则称为相关学习规则。这种规则只根据连接间的激活水平改变权系数。常用于自联想网络, 如 Hopfield 网络。

第二类学习规则称为纠错学习规则。这种规则根据输出节点的外部反馈改变权系数。

在方法上它和梯度下降法等效,按局部改善最大的方向一步步进行优化,从而最终找到全局优化值。感知器学习就采用这种纠错学习规则,如BP算法。用于统计性算法的模拟退火算法也属于这种学习规则。

第三类学习规则称为无教师学习规则。它是一种对输入检测进行自适应的学习规则。ART网络的自组织学习算法即属于这一类。

22.1.6 SPSS 神经网络模型

使用SPSS神经网络,可以帮助探索数据中微妙或者隐藏的模式。这个附加模块可以帮助发现数据中更复杂的关系,产生更有效果的预测模型。SPSS神经网络是对SPSS Statistics Base,以及附加模块中传统统计方法的一个补充。可以使用SPSS神经网络发现数据中间的新关系,然后用传统的统计技术检验其显著性。

1. 为什么要使用神经网络

神经网络是一个非线性的数据建模工具集合,它包括输入层和输出层、一个或者多个隐藏层。神经元之间的连接赋予相关的权重,训练算法在迭代过程中不断调整这些权重,从而使得预测误差最小化并给出预测精度。可以设置网络的训练条件,从而控制训练的停止条件,以及网络结构,或者让算法自动选择最优的网络结构。在许多领域,都可以将SPSS神经网络和其他的统计分析过程结合起来,获得更深入、清晰的洞察力。例如,在市场研究领域,可以建立客户档案发现客户的偏好;在数据库营销领域,可以进行客户细分,优化市场活动的响应。

在金融分析方面,可以使用SPSS神经网络分析申请人的信用状况,探测可能的欺诈。在运营分析方面,也可以使用这个新工具管理现金流、优化供应链。此外,在科学和医疗方面的应用包括预测医疗费用、医疗结果分析、预测住院时间等。

2. SPSS 中的神经网络过程

SPSS神经网络,包括多层感知器(MLP)或者RBF两种方法。

这两种方法都是有监督的学习技术——也就是说,它们根据输入的数据映射出关系。这两种方法都采用前馈结构,意思是数据从一个方向进入,通过输入节点、隐藏层最后进入输出节点。对过程的选择受到输入数据的类型和网络的复杂程度的影响。此外,多层感知器可以发现更复杂的关系,RBF的速度更快。MLP可以发现更复杂的关系,而通常来说RBF更快。

使用这两种方法的任何一种,可以将数据拆分成训练集、检验集、验证集三种。训练集用来估计网络参数。检验集用来防止过度训练。验证样本用来单独评估最终的神经网络,它将应用于整个数据集和新数据。设置的因变量可以是连续型、分类型或者两者的组合。如果因变量是连续型,神经网络预测的连续值是近似于输入数据的某个连续函数的“真实”值。如果因变量是分类型,神经网络会根据输入数据,将记录划分为最适合的类别。

可以通过选择分析中拆分数据集,网络结构的排序,计算方法等调整神经网络程序。最后,可以通过图形或者表格,在当前活动的数据集中保存可选的临时变量,并且将模型导出成XML格式对新的数据进行拆分。

3. SPSS 神经网络的特性

(1) MLP (多层感知器)

MLP 通过多层感知器来拟合神经网络。多层感知器是一个前馈式有监督的结构。它可以包含多个隐藏层。一个或者多个因变量, 这些因变量可以是连续型、分类型或者两者的结合。如果因变量是连续型, 神经网络预测的连续值是输入数据的某个连续函数。如果因变量是分类型, 神经网络会根据输入数据, 将记录划分为最适合的类别。

预测。

- 因子。
- 协变量。

选项“除外”列出 MLP 中需要排除的因子或者协变量。当因子或者协变量包含大量的变量时, 这个选项很有用。

选项“缩放”对协变量和因变量进行变换。

- 因变量(如果需要变换): 标准化, 正态化, 调整的正态化, 或者无。
- 协变量: 标准化, 正态化, 调整的正态化或者无。

选项“拆分”用来设定对当前活动数据集的拆分方法。训练样本用来训练神经网络、检验集是一个独立的数据集, 用来跟踪预测无法来防止过度训练。验证集是另外一个独立的数据集, 用来评估最后的神经网络。可以如下设置。

- 相对记录数来随机分配训练样本。
- 相对记录数来随机分配检验样本。
- 相对记录数来随机分配验证样本。
- 使用变量对样本进行拆分。

选项“结构”用来设置神经网络的结构, 可以如下设置。

- 是否使用自动选择结构。
- 神经网络的隐藏层个数。
- 隐藏层单元之间的激活函数(双曲函数或者 S 型函数)。
- 输出层单元之间的激活函数(标识, 双曲, S 型, SoftMax 函数)。

选项“标准”设定 MLP 的计算参数。例如, 训练类型决定了神经网络如何处理训练数据, 包括批处理训练、在线训练、小批量训练, 可以如下设置。

- 每个小批量的训练记录数。
- 当选择自动化结构或者小批量训练模式时, 内存中存储的最大记录个数。
- 优化算法决定突触权重: 梯度下降法、共轭梯度下降法。
- 梯度下降优化方法的初始学习率。
- 当使用在线或者小批量训练模式时, 梯度下降的学习率下限。
- 梯度下降优化算法的动量率。
- 共轭梯度下降法的初始 lambda。
- 共轭梯度下降法的 sigma。
- 初始权重区间 $[a_0 - a, a_0 + a]$ 。

选项“停止训练”决定神经网络停止训练的规则, 可以如下设置。

- 预测误差下降的次数。
- 训练时间或者最大训练时间。
- 最大收敛次数。
- 训练误差的相对变化率。
- 训练误差率准则。

选项“缺失”用来控制分类变量(因子和分类因变量)的缺失值是否被作为有效值使用。

选项“打印”指定输出内容,也可以请求一个敏感性分析,可以如下设置。

- 处理过程设置概要。
- 神经网络的基本信息,包括因变量、输入和输出单元个数、隐藏层单元个数、激活函数。
- 神经网络输出结果的概要信息,包括总体平均误差、停止规则、训练时间。
- 每个分类因变量的分类表突触权重,也就是连接第 $i-1$ 层第 j 个单元和第 i 层第 k 个单元的系数估计值。
- 敏感性分析,用来计算每个预测元对神经网络的影响的重要性。

选项“画图”指定输出的图形,可以进行如下选择。

- 网络图。
- 每个因变量的预测值和观测值图。
- 连续型因变量的残差图。
- 分类变量的 ROC 曲线。
- 分类因变量的累积收益图。
- 分类因变量的提升图。

⑪ 选项“保存”可以将生成的临时变量保存到当前数据集中,可以进行如下保存。

- 预测值或者分类。
- 预测的伪概率。

⑫ 选项“输出文件”将神经网络的结构输出保存成包含突触权重的 XML 格式。

(2) RBF

RBF 程序拟和一个前馈型、有监督学习的 RBF 网络,包括输入层、隐藏层(径向基函数层)、输出层。输入矢量通过隐藏层传递到 RBF,类似 MLP, RBF 可以进行预测和分类。

RBF 程序分以下两个阶段训练网络。

- 程序通过聚类方法确定 RBF,以及每个 RBF 的中心和宽度。
- 估计 RBF 的连接权重。在预测和分类中都使用激活函数作为均方误差函数,使用普通最小二乘法求均方误差的最小值。

由于 RBF 训练过程分两个阶段,因此,一般情况下, RBF 网络的训练速度优于 MLP。MLP 和 RBF 的选项基本相同,除了以下几个方面。

- 如果使用“结构”选项,用户可以指定隐藏层的高斯径向基函数:标准 RBF 或者普通 RBF。
- 当使用“准则”选项时,用户可以指定 RBF 的计算参数,指定隐藏层单元的交叠方式。

22.2 SPSS 神经网络模型的设置

22.2.1 多层感知器（MLP）分析过程的参数设置

选择菜单“分析（Analyze）神经网络（Neural Networks）多层感知器（Multilayer Perceptron）”，则弹出如图 22-3 所示的对话框，此对话框用于设置多层感知器的各种参数。此界面中有 8 个标签，即变量（Variables）、分区（Partitions）、体系结构（Architecture）、训练（Training）、输出（Output）、保存（Save）、导出（Export）、选项（Options）。

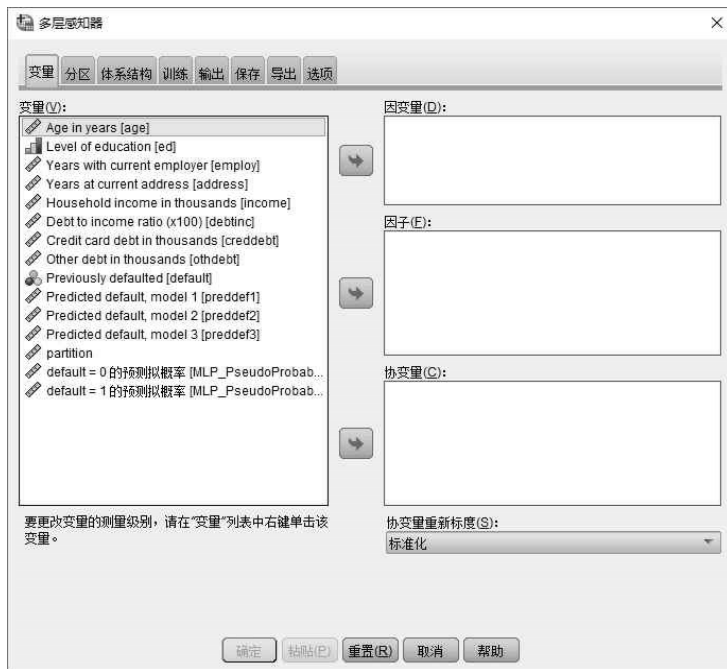


图 22-3 “多层感知器（Multilayer Perceptron）设置”对话框

1. 变量（Variables）

图 22-3 中的左边是待分析的列表框，其他各个选项框含义如下。

- 因变量（Dependent Variables）：用于选入因变量。
- 因子（Factors）：用于选入因变量。
- 协变量（Covariates）：用于设置协变量。
- 协变量重新标度（Rescaling of Covariates）：用于设置协变量的标度，其下拉菜单中有四个选择项，分别是标准化（Standardized）、正态化（Normalized）、调整后正态化（Adjusted Normalized）以及无（None）。

2. 分区（Partitions）

单击图 22-3 中的“分区（Partitions）”标签，则弹出如图 22-4 所示的对话框，此界面

组成部分如下。

变量 (Variables) : 用于存放待分析的变量。

分区数据集 (Partition Dataset) : 用于设置分区数据集, 有两个选项。

- 根据个案的相对数量随机分配个案 (Randomly assign cases based on relative number of cases) : 根据个案的相对数量随机的分配个案, 其下的选项栏有分区 (Partition)、相对数目 (Relative Number)、训练 (Training)、检验 (Test)、支持 (Holdout)、总计 (Total)。
- 使用分区变量来分配个案 (Use portioning variable to assign cases) : 使用分区变量分配个案, 其下的分区变量 (Partitioning Variable) 表示分区变量。



图 22-4 “分区 (Partitions) 设置”对话框

3. 体系结构 (Architecture)

单击图 22-4 中的“体系结构 (Architecture)”标签, 则弹出如图 22-5 所示的对话框, 各个组成部分如下所述。

(1) 体系结构自动选择 (Automatic architecture selection)

- 隐藏层中的最小单元数 (Minimum Number of Units in Hidden Layer)
- 隐藏层中的最大单元数 (Maximum Number of Units in Hidden Layer)

(2) 定制体系结构 (Custom Architecture)

隐藏层 (Hidden Layers)

- 一层。
- 两层。

激活函数 (Activation Function)

- 双曲正切 (Hyberbolic Tangent)
- S 型函数 (Sigmoid)
- 单元数 (Number of Units)
- 自动计算 (Automatically Compute)
- 定制 (Custom) : 用户自定义 , 如下有隐藏层 1 (Hidden Layer1) , 隐藏层 2 (Hidden Layer2)

(3) 输出层 (Output Layer)

激活函数 (Activation Function)

- 恒等函数 (Identity)
- Softmax。
- 双曲正切 (Hyberbolic Tangent)
- S 型。
- 标度因变量重新标度 (Rescaling of Scale Dependent Variables)
- 标准化 (Z)(Standardized)
- 正态化 (Normalized)
- 调整后正态化 (Adjusted Normalized)
- 无 (None)



图 22-5 “体系结构 (Architecture) 设置”对话框

4. 训练 (Training)

单击图 22-5 中的“训练 (Training)”标签, 则弹出如图 22-6 所示的对话框。

训练类型 (Type of Training)

- 批次 (Batch)
- 联机 (Online)
- 小批次 (Mini-batch), 选择此项后激活其下的选项栏, 自动计算 (Automatically compute) 表示自动计算; 定制 (Custom) 表示拥护自定义, 在其下的记录数 (Number of Records) 选项栏中填入记录数。
- 优化算法 (Optimization Algorithm)
- 标度共轭梯度法 (Scaled Conjugate Gradient)
- 梯度下降法 (Gradient Descent)
- 训练选项 (Training Options)
- 初始 Lambda 值 (Initial Lambda)
- 初始的 Sigma 值 (Initial Sigma)
- 区间中心点 (Interval Center)
- 区间偏移量 (Interval Offset)



图 22-6 “训练 (Training) 设置”对话框

5. 输出 (Output)

单击图 22-6 中的“输出 (Output)”标签, 则弹出如图 22-7 所示的对话框, 各个部分组成如下。

网络结构 (Network Structure)

- 描述 (Description)
- 图 (Diagram)
- 突触权重 (Synaptic Weights)

网络性能 (Network Performance)、

- 模型摘要 (Model Summary)、
- 分类结果 (Classification Results)、
- ROC 曲线 (ROC Curve)、
- 累积增益图 (Cumulative Gains Chart)、
- 效益图 (Lift Chart)、
- 预测—实测值 (Predicted by Observed Chart)、
- 残差—预测图 (Residual by Predicted Chart)、

个案处理摘要 (Cases Processing Summary)、

自变量重要性的分析 (Independent Variable Importance Analysis)、



图 22-7 “输出 (Output) 设置”对话框

6. 保存 (Save)

单击图 22-7 中的“保存 (Save)”标签，则弹出如图 22-8 所示的对话框。

保存每个因变量的预测值或类别 (Save Predicted Value or Category for Each Dependent Variable)、

保存每个因变量预测拟概率 (Save Predicted Pseudo-probability for Each Dependent Variable)、

保存变量的名称 (Names of Saved Variables) 选项栏：用于保存变量名称。

- 自动生成唯一名称 (Automatically Generate Unique Names)：自动生成唯一的名称。如果要在每次运行模型时将一组新的保存变量添加到数据集，则选择此项。
- 定制名称 (Custom Names)：为变量指定名称，如果选择此项，具有相同名称或者根名称的任何现有自变量将在每次运行模型时刻被替换。



图 22-8 “保存 (Save) 设置”对话框

7. 导出 (Export)

单击图 22-8 中的“导出 (Export)”标签，则弹出如图 22-9 所示的对话框。此对话框用于导出数据。其中将突触权重估算值导出至 XML 文件 (Export synaptic weight estimates to XML file) 选项表示将突触权重估算数据导出到 XML 文件之中。

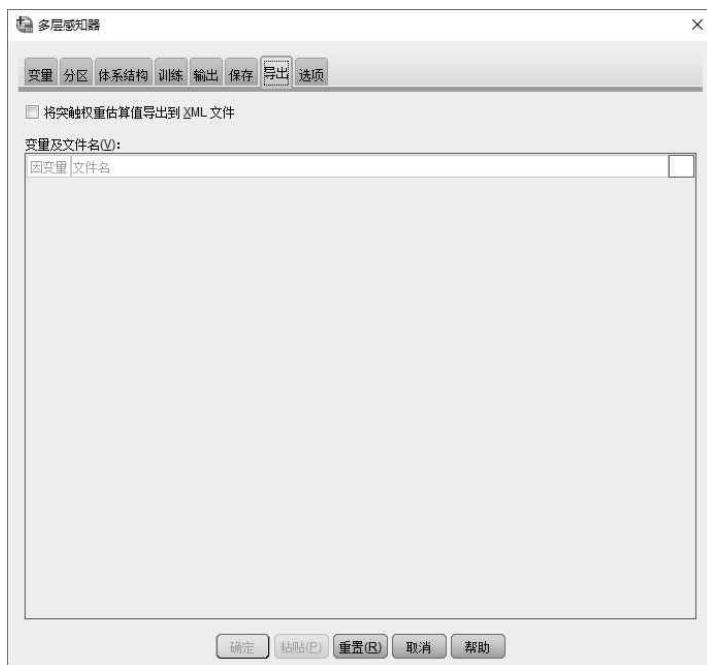


图 22-9 “导出 (Export) 设置”对话框

8. 选项 (Options)

单击图 22-9 中的“选项 (Options)”标签, 则弹出如图 22-10 所示的对话框, 各部分组成如下。

用户缺失值 (User-Missing Values)

- 指定如何处理在因子及分类因变量中具有用户缺失值的个案 (Specify how to treat cases with user-missing values on factors and categorical dependent variables), 其下有两个选项即排除 (Exclude) 和包括 (Include)。
- 始终排除在协变量或标度因变量中具有用户缺失值的个案 (Cases with user-missing values on covariates or scale dependent variables are always excluded)。

中止规则 (Stopping Rules)

- 误差未减少情况下的最大步骤数 (Maximum Steps without a Decrease in Error)。
- 用于计算预测误差的数据 (Data to Use for Computing Prediction Error), 其下有自动选择 (Choose Automatically); 训练及检验数据 (Both Training and Test Data)。
- 最长训练时间 (Maximum Training Time), 在其后的 Minutes 中填写数据。
- 最长训练时程 (Maximum Training Epochs), 其下有自动计算 (Compute Automatically) 指定自定义值 (Specify Custom Value), 其后的 Maximum number of epochs 表示最大时程数。
- 训练误差的最小相对变化 (Minimum Relative Change in Training Error)。
- 训练误差率的最小相对变化 (Minimum Relative Change in Training Error Ratio)。
- 存储在内存中的最大个案数 (Maximum Cases to Store in Memory)。



图 22-10 “选项 (Options) 设置”对话框

22.2.2 径向基函数(RBF)分析过程的参数设置

RBF 的设置和多层感知器的设置基本相同,只是 RBF 不需要设置 Training 选项,即不需要训练数据集。选择菜单“分析(Analyze) 神经网络(Neural Networks) 径向基函数(Radial Basis Function)”,则弹出如图 22-11 所示的对话框,此框用于设置 RBF 的各种参数。此界面中共有 7 个标签,即 Variables(变量)、Partitions(分区)、Architecture(体系结构)、Output(输出)、Save(保存)、Export(导出)、Options(选项)。下面介绍 RBF 设置的不同之处,包括标签 Architecture、Option。

1. 变量设置

图 22-11 中的变量选项栏与多层感知器的设置基本相同,只是在因变量(Dependent Variables)选项栏的下方需要设置标度因变量重新标度(Rescaling of Scale Dependent Variables),下拉菜单有四个选项,分别是标准化(Standardized)、正态化(Normalized)、调整后正态化(Adjusted Normalized),以及无(None)。

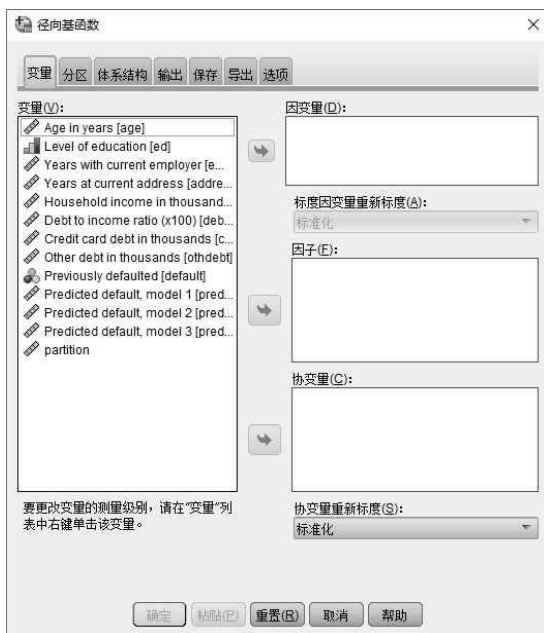


图 22-11 “径向基函数(RBF)的设置”对话框

2. 体系结构(Architecture)设置

单击图 22-11 中的“体系结构(Architecture)”标签,则弹出如图 22-12 所示的对话框,此对话框用于设置体系结构,各个组成部分如下。

隐藏层中的单元数(Number of Units in Hidden Layer)。

- 在某个范围内查找最佳单元数(Find the Best Number of Units within a Range)。其下选项栏用于设置范围,有自动计算范围(Automatically Compute Range);使用指定范围(Use a Specified Range),其下的选项栏有最小值(Minimum),最大值(Maximum)。

- 使用指定单元数 (Use a Specified Number of Units), 其下的数值 (Number) 选项栏用于指定单元数量。
- 隐藏层激活函数 (Activation Function for Hidden Layer)
- 正态化径向基函数 (Normalized Radial Basis Function)
- 普通径向基函数 (Ordinary Radial Basis Function)
- 隐藏单位之间的重叠 (Overlap Among Hidden Units)
- 自动计算允许的重叠量 (Automatically Compute the Amount of Overlap to Allow)
- 允许指定数量的重叠量 (Allow Specified Amount of Overlap), 在其下的重叠因子 (Overlapping Factor) 选项栏中填入重叠因子。



图 22-12 “体系结构 (Architecture) 设置”对话框

3. 选项 (Option) 设置

单击图 22-11 中的“选项 (Option)”标签, 则弹出如图 22-13 所示的对话框各部分组成如下。

用户缺失值 (User-Missing Values)

指定如何处理在因子及分类因变量中具有用户缺失值的个案 (Specify how to treat cases with user-missing values on factors and categorical dependent variables), 其下有两个选项。

- 排除 (Exclude)
- 包括 (Include)

始终排除在协变量或标度因变量中具有用户缺失值的个案 (Cases with user-missing values on covariates or scale dependent variables are always excluded)

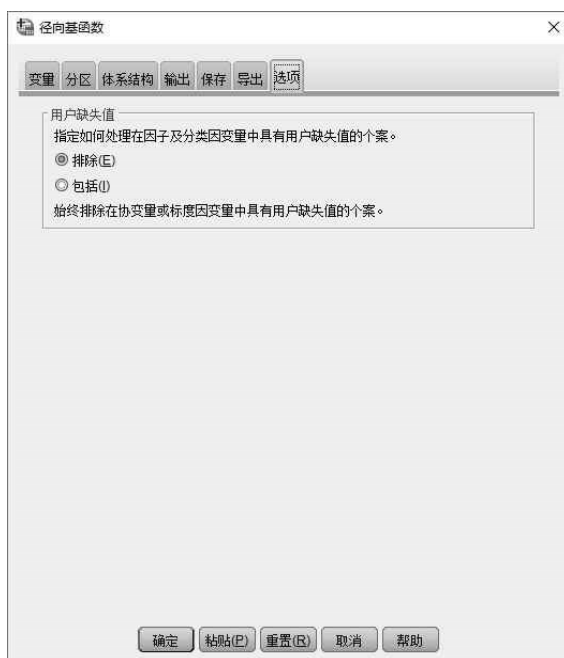


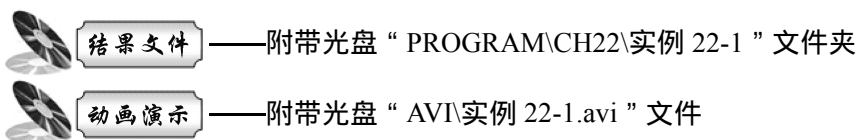
图 22-13 “选项 (Options) 设置”对话框

22.3 实例分析

本节中将利用 SPSS 中的神经网络模块来进行分析，采用的数据集为 SPSS 自带的数据集 bankloan.sav，此数据在前面的章节中已经介绍过，数据集共包含 12 个变量，包括 age（年龄）、ed（学历水平）、address（当前地址）、income（收入）等，数据集包含 850 个调查样本。数据集的格式如图 22-14 所示，其中 ed 变量共分为 4 个范围，即 Did not complete high school、High school degree、Some college、College degree，以及 Post-undergraduate degree。变量 default 表示违约，有两个选择项，即 No 和 Yes。下面就利用此数据集中的 700 个样本数据作为训练数据集来创建一个多层感知器的神经网络模型，并利用创建的模型来分析余下的 150 个调查用户的信用记录，以观察这 150 个用户的信用是好还是坏。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|----|----------|----|----|------|-------------------------------------|------------|----|----|----|----|----|
| 1 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 4 | 右 | 标度 | 输入 |
| 2 | ed | 数字 | 4 | 0 | Level of education {1, Did not c... | 无 | 无 | 4 | 右 | 有序 | 输入 |
| 3 | employ | 数字 | 4 | 0 | Years with curr... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 4 | address | 数字 | 4 | 0 | Years at curren... | 无 | 无 | 7 | 右 | 标度 | 输入 |
| 5 | income | 数字 | 8 | 2 | Household inco... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 6 | debtinc | 数字 | 8 | 2 | Debt to income... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 7 | creddebt | 数字 | 8 | 2 | Credit card deb... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 8 | othdebt | 数字 | 8 | 2 | Other debt in th... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 9 | default | 数字 | 4 | 0 | Previously defa... | {0, No}... | 无 | 7 | 右 | 名义 | 输入 |
| 10 | preddef1 | 数字 | 11 | 5 | Predicted defau... | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 11 | preddef2 | 数字 | 11 | 5 | Predicted defau... | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 12 | preddef3 | 数字 | 11 | 5 | Predicted defau... | 无 | 无 | 11 | 右 | 标度 | 输入 |

图 22-14 数据集 bankloan.sav 的格式



22.3.1 参数设置

首先产生随机数来选择样本数据集，选择菜单“转换（Transform） 随机数字生成器（Random Number Generators）”，则弹出如图 22-15 所示的对话框，选择“设置起点（Set Starting Point）”选项栏，并选中“固定值（Fixed Value）”选项，填入 9191972，然后单击图 22-15 中的“确定”按钮。

然后选择菜单“转换（Transform） 计算变量（Compute Variable）”，则弹出如图 22-16 所示的对话框，在“目标变量（Target Variable）”选项栏中填入变量名 partition，然后在“数学表达式（Numeric Expression）”选项框中填入计算表达式 $2 * RV.BERNOULLI(0.7) - 1$ ，此公式用于产生 bernoulli 分布数据，数据集的名称为 partition。

设置完成后单击界面“确定”按钮进行计算。

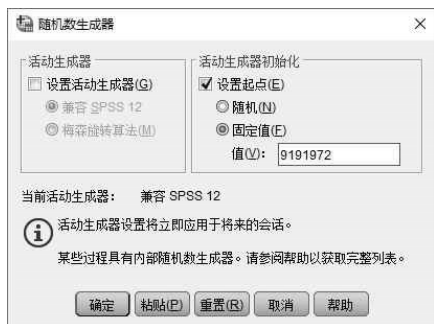


图 22-15 “随机数字生成器（Random Number Generators）设置”对话框



图 22-16 “计算变量（Compute Variable）设置”对话框

生成随机数以后，则选择菜单“分析（Analyze） 神经网络（Neural Networks） 多层感知器（Multilayer Perceptron）”，则弹出如图 22-17 所示的对话框。选择变量 Previously defaulted [default] 到“因变量（Dependent Variables）”选项栏中。选择变量 Level of education [ed] 到“因子（Factors）”选项栏中。选择变量 age、employ、address、income、debtinc、creddebt、othdebt 到“协变量（Covariates）”选项栏中。

然后单击“分区（Partitions）”标签，弹出如图 22-18 所示对话框，选中“使用分区变量分配个案（Use Partitioning Variable to Assign Cases）”选项栏，然后选中变量 partition 到“分区变量（Partitioning Variable）”选项栏中。

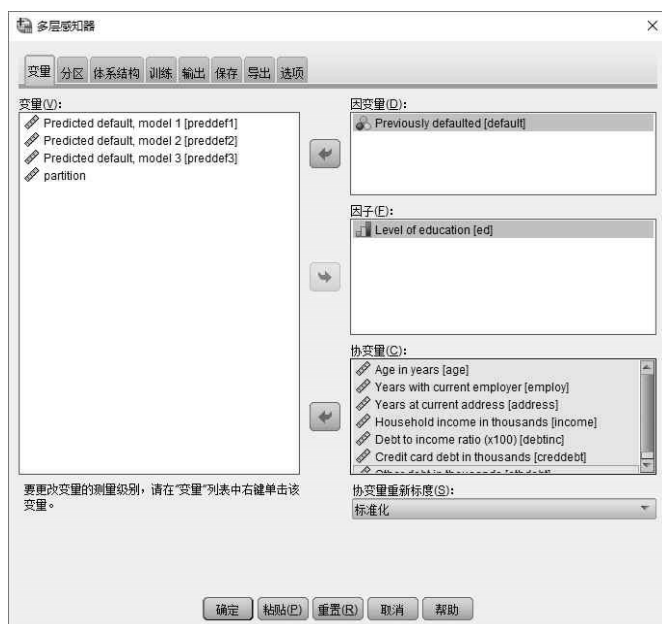


图 22-17 “多层感知器 (Multilayer Perceptron) 变量设置”对话框



图 22-18 “分区 (Partitions) 设置”对话框

然后单击“输出 (Output)”标签,弹出如图 22-19 所示对话框。选择“ROC 曲线 (ROC Curve)”、“累积增益图 (Cumulative Gains Chart)”、“效益图 (Lift Chart)”,以及“预测-实测图 (Predicted by Observed Chart)”选项栏,去掉“图表 (Diagram)”选项。最后选择“自变量重要性分析 (Independent Variable Importance Analysis)”选项栏,然后单击“确定”按钮进行分析。



图 22-19 “输出（Output）设置”对话框

22.3.2 结果分析

设置完成以后则单击主界面中的“确定”按钮进行分析，首先输出的是案例处理信息汇总，如图 22-20 所示。包括 700 个有效样本数据，是从 850 个总体中选出的，然后排除 150 个样本数据。

| 个案处理摘要 | | | |
|--------|----|-----|--------|
| | | 个案数 | 百分比 |
| 样本 | 训练 | 499 | 71.3% |
| | 坚持 | 201 | 28.7% |
| 有效 | | 700 | 100.0% |
| 排除 | | 150 | |
| 总计 | | 850 | |

图 22-20 案例处理信息汇总

然后是网络信息，如图 22-21 所示，包括输入层（Input Layer）、隐藏层（Hidden Layer），以及输出层（Output Layer），卡尔变量有 7 个，隐藏层的个数为 1 个，包含 4 个单元。

接着输出的是模型信息，如图 22-22 所示，给出了训练数据集的各种信息。图 22-23 是回判分析表格，给出了样本数据集中的回判正确率等信息。如训练集中变量 No 所对应的样本数为 375 个，有 347 判为 No，28 个样本判为 Yes，所以回判正确率为 92.5%，同样变量 Yes 对应的回判正确率为 59.7%，看来效果不是很好，平均回判正确率为 84.4%。而对 Holdout 样本数据集，Yes 对应的样本，其回判正确率仅为 45.8%，也同样不是很好。

| 网络信息 | | | | |
|------|--------------------------|---|-------------------------------|---------|
| 输入层 | 因子 | 1 | Level of education | |
| | 协变量 | 1 | Age in years | |
| | | 2 | Years with current employer | |
| | | 3 | Years at current address | |
| | | 4 | Household income in thousands | |
| | | 5 | Debt to income ratio (x100) | |
| | | 6 | Credit card debt in thousands | |
| | | 7 | Other debt in thousands | |
| | 单元数 ^a | | | 12 |
| | 协变量的重新标度方法 | | | 标准化 |
| 隐藏层 | 隐藏层数 | | | 1 |
| | 隐藏层 1 中的单元数 ^a | | | 4 |
| | 激活函数 | | | 双曲正切 |
| | 误差函数 | | | 交叉熵 |
| 输出层 | 因变量 | 1 | Previously defaulted | |
| | 单元数 | | | 2 |
| | 激活函数 | | | Softmax |
| | 误差函数 | | | 交叉熵 |

a. 排除偏差单元

图 22-21 网络信息

| 模型摘要 | | |
|------|---------------------------|---------------|
| 训练 | 交叉熵误差 | 156.605 |
| | 不正确预测百分比 | 15.6% |
| | 使用的中止规则 | 超出最大时程数 (100) |
| | 训练时间 | 0:00:00.28 |
| 坚持 | 不正确预测百分比 | 25.4% |
| | 因变量: Previously defaulted | |

图 22-22 模型信息

| 分类 | | | | |
|----|-------|-------|-------|-------|
| 样本 | 实测 | 预测 | | 正确百分比 |
| | | No | Yes | |
| 训练 | No | 347 | 28 | 92.5% |
| | Yes | 50 | 74 | 59.7% |
| | 总体百分比 | 79.6% | 20.4% | 84.4% |
| 坚持 | No | 123 | 19 | 86.6% |
| | Yes | 32 | 27 | 45.8% |
| | 总体百分比 | 77.1% | 22.9% | 74.6% |

因变量: Previously defaulted

图 22-23 判别分析表

基于上述所训练的模型结果不是很好，所以要重新进行设置，选择菜单“转换 (Transform) 计算变量 (Compute Variable)”，则弹出如图 22-24 所示的对话框，在“目标变量 (Target Variable)”选项栏中填入变量名 partition，然后在“数值表达式 (Numeric Expression)”选项框中重新填入计算表达式 $\text{partition} - \text{Rv.BERNOULLI}(0.2)$ ，此公式用于产生 bernoulli 分布数据，数据集的名称同样为 partition。



图 22-24 “计算变量 (Compute Variable) 设置”对话框

然后单击图 22-24 中的“如果 (If)”按钮进行条件设置，如图 22-25 所示。选择“如果个案满足条件则包括 (Include if Case Satisfies Condition)”选项栏，并输入 $\text{partition} > 0$ ，然后单击“继续”按钮。



图 22-25 “如果 (If) 设置”对话框

设置好后则选择图 22-17 中的“保存 (Save)”标签，则弹出如图 22-26 所示对话框，选中“保存各因变量的预测拟概率或类别 (Save Predicted Pseudo-probability for Each Dependent Variable)”选项，然后单击“确定”按钮进行分析。

运行的结果如下，首先还是案例处理信息汇总，如图 22-27 所示。共有 499 个初始样本，其中 401 个用于训练，98 个用于进行检验。



图 22-26 “保存 (Save) 设置”对话框

| 个案处理摘要 | | | |
|--------|----|-----|--------|
| | | 个案数 | 百分比 |
| 样本 | 训练 | 401 | 57.3% |
| | 检验 | 98 | 14.0% |
| | 坚持 | 201 | 28.7% |
| 有效 | | 700 | 100.0% |
| 排除 | | 150 | |
| 总计 | | 850 | |

图 22-27 案例信息汇总

接着输出的是如图 22-28 所示的网络信息输出。与上面的网络信息不同的是隐藏层的个数 (Number of Units in Hidden Layer) 为 7 个。

| 网络信息 | | |
|------------------|--------------------------|---------------------------------|
| 输入层: | 因子 | 1 Level of education |
| | 协变量 | 1 Age in years |
| | | 2 Years with current employer |
| | | 3 Years at current address |
| | | 4 Household income in thousands |
| | | 5 Debt to income ratio (x100) |
| | | 6 Credit card debt in thousands |
| | | 7 Other debt in thousands |
| 单元数 ^a | | 12 |
| 协变量的重新标度方法 | | 标准化 |
| 隐藏层: | 隐藏层数 | 1 |
| | 隐藏层 1 中的单元数 ^a | 7 |
| | 激活函数 | 双曲正切 |
| 输出层: | 因变量 | 1 Previously defaulted |
| | 单元数 | 2 |
| | 激活函数 | Softmax |
| | 误差函数 | 交叉熵 |

a. 排除偏差单元

图 22-28 网络信息输出

如图 22-29 所示的是模型的信息，包括正确判断的百分比，训练（Training）为 18.7%，检验（Testing）为 19.4%，保持（Holdout）为 20.9%。

| 模型摘要 | | | |
|---------------------------|----------|-------------------------------|------------|
| 训练 | 交叉熵误差 | | 159.674 |
| | 不正确预测百分比 | | 18.7% |
| | 使用的中止规则 | 误差在 1 个连续步骤中没有减小 ^a | |
| | 训练时间 | | 0:00:00.19 |
| 检验 | 交叉熵误差 | | 40.972 |
| | 不正确预测百分比 | | 19.4% |
| 坚持 | 不正确预测百分比 | | 20.9% |
| 因变量: Previously defaulted | | | |
| a. 误差计算基于检验样本。 | | | |

图 22-29 模型输出信息

图 22-30 输出的是判别分类结果，给出了个类别的判别正确率。

| 分类 | | | | |
|---------------------------|-------|-------|-------|-------|
| 样本 | 实测 | 预测 | | 正确百分比 |
| | | No | Yes | |
| 训练 | No | 276 | 33 | 89.3% |
| | Yes | 42 | 50 | 54.3% |
| | 总体百分比 | 79.3% | 20.7% | 81.3% |
| 检验 | No | 56 | 10 | 84.8% |
| | Yes | 9 | 23 | 71.9% |
| | 总体百分比 | 66.3% | 33.7% | 80.6% |
| 坚持 | No | 123 | 19 | 86.6% |
| | Yes | 23 | 36 | 61.0% |
| | 总体百分比 | 72.6% | 27.4% | 79.1% |
| 因变量: Previously defaulted | | | | |

图 22-30 判别分类结果

然后输出的是 ROC 曲线，如图 22-31 所示，ROC 曲线是二元判决中用来比较判别方法优劣的一种曲线，通常所说一种判别方法优于另外一种判别方法，从 ROC 曲线上来说，就是曲线的横坐标尽可能的小而纵坐标尽可能的大，即较好的方法的 ROC 曲线应该始终在较差丰富的 ROV 曲线的左上方。如图 22-32 所示的是增益图形。

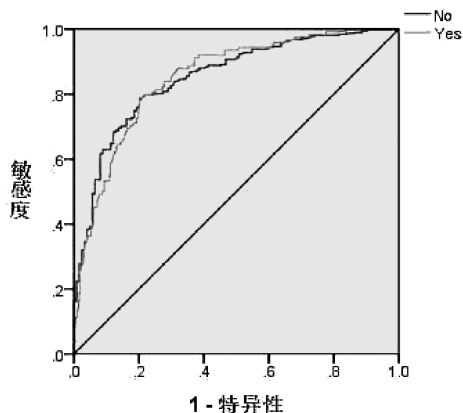


图 22-31 ROC 曲线

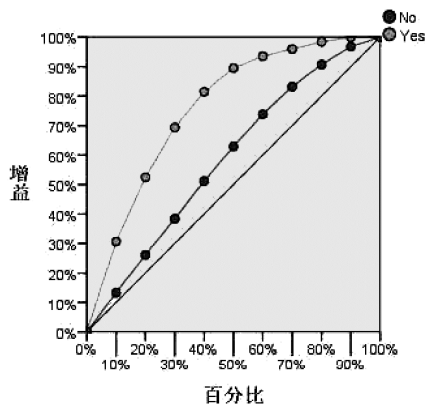
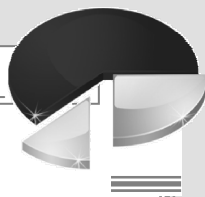


图 22-32 增益图形



第 23 章 信用风险分析

信用风险 (Credit Risk) 又称违约风险, 是指交易对手未能履行约定契约中的义务而造成经济损失的风险, 即受信人不能履行还本付息的责任而使授信人的预期收益与实际收益发生偏离的可能性, 它是金融风险的主要类型。

本章将利用 SPSS 软件来分析信用风险模型, 例如, 多元逻辑模型、判别分析模型, 以及分类树模型。



本讲内容

- 信用风险概述
- 实例分析

23.1 信用风险概述

23.1.1 信用风险基本概念

信用风险是借款人因各种原因未能及时、足额偿还债务或银行贷款而违约的可能性。发生违约时, 债权人或银行必将因为未能得到预期的收益而承担财务上的损失。信用风险是由下面两方面的原因造成的。

一是经济运行的周期性。

在处于经济扩张期时, 信用风险降低, 因为较强的赢利能力使总体违约率降低。在处于经济紧缩期时, 信用风险增加, 因为赢利情况总体恶化, 借款人因各种原因不能及时足额还款的可能性增加。

二是对于公司经营有影响的特殊事件的发生。

这种特殊事件发生与经济运行周期无关, 并且对公司经营有重要的影响。例如, 产品的质量诉讼。举具体事例来说, 当人们知道石棉对人类健康有影响的事实时, 所发生的产品的责任诉讼是 Johns-Manville 公司, 一个著名的在石棉行业中处于领头羊位置的公司破产并无法偿还其债务。

信用风险有以下四个主要特征。

- 客观性，不以人的意志为转移。
- 传染性，一个或少数信用主体经营困难或破产就会导致信用链条的中断和整个信用秩序的紊乱。
- 可控性，其风险可以通过控制降到最低。
- 周期性，信用扩张与收缩交替出现。

由于信用风险会对公司或个人的利益产生很大的影响，因此信用风险管理变成很重要的工作，较大的公司常有专门人员，针对各个交易对象的信用状况作评估来衡量可能的损益，以及减低可能的损失。

对于商业银行，当借款人对银行贷款违约时，商业银行是信用风险的承受者。银行因为两个原因会受到相对较高的信用风险。首先，银行的放款通常在地域上和行业上较为集中，这就限制了通过分散贷款而降低信用风险的方法的使用。其次，信用风险是贷款中的主要风险。随着无风险利率的变化，大多数商业贷款都设计成是浮动利率的。这样，无违约利率变动对商业银行基本上没有什么风险。而当贷款合约签定后，信用风险贴水则是固定的。如果信用风险贴水升高，则银行就会因为贷款收益不能弥补较高的风险而受到损失。

管理信用风险有多种方法，下面进行具体介绍。

23.1.2 信用风险度量方法

信用风险度量方法有很多种，包括传统的信用风险评价、多变量信用风险判别模型，以及现代金融工程模型，首先介绍传统的信用风险评价。

1. 传统的信用风险评价方法

(1) 要素分析法

要素分析法是通过定性分析有关指标来评价客户信用风险时所采用的专家分析法。常用的要素分析法是 5C 要素分析法，它主要集中在借款人的道德品质（Character）、还款能力（Capacity）、资本实力（Capital）、担保（Collateral）和经营环境条件（Condition）5 个方面进行全面的定性分析，以判别借款人的还款意愿和还款能力。

根据不同的角度，有的将分析要素归纳为“5W”因素，即借款人（Who）、借款用途（Why）、还款期限（When）、担保物（What）及如何还款（How）。还有的归纳为“5P”因素，即个人因素（Personal）、借款目的（Purpose）、偿还（Payment）、保障（Protection）和前景（Perspective）。无论是“5C”、“5W”还是“5P”，其共同之处都是先选取一定特征目标要素，然后对每一要素评分，使信用数量化，从而确定其信用等级，以其作为其销售、贷款等行为的标准和随后跟踪监测期间的政策调整依据。

(2) 特征分析法

特征分析法是目前在国外企业信用管理工作中应用较为普遍的一种信用分析工具。它是从客户的种种特征中选择出对信用分析意义最大、直接与客户信用状况相联系的若干因素，将其编为几组，分别对这些因素评分并综合分析，最后得到一个较为全面的分析结果。一般所分析的特征包括客户自身特征、客户优先性特征、信用及财务特征等。特征分析法的主要用途是对客户的资信状况作出综合性的评价，它涵盖了反映客户经营实力和发

展潜力的所有重要指标,这种信用风险分析方法主要由信用调查机构和企业内部信用管理部门使用。

(3) 财务比率分析法

信用风险往往是由财务危机导致的,因此,可以通过及早发现和找出一些特征财务指标,判断评价对象的财务状况和确定其信用等级,从而为信贷和投资提供决策依据。财务比率综合分析法就是将各项财务分析指标作为一个整体,系统、全面、综合地对企业财务状况和经营情况进行剖析、解释和评价。这类方法的主要代表有杜邦财务分析体系和沃尔比重评分法。杜邦财务分析体系是由美国杜邦公司创立的,它以净值报酬率为龙头,以资产净利润率为核心,重点揭示企业获利能力及其前因后果,通过对某项综合性较强的财务比率的逐层分解,将相关财务指标联系起来,形成一个综合体系,以便清楚地反映各项财务指标的相互关系。沃尔比重评分法是由财务综合评价领域的著名先驱者之一亚历山大·沃尔创立的,他把若干个财务比率用线性关系结合起来,以此评价企业的信用水平。他选择了7种财务比率,即流动比率、产权比率、固定资产比率、存货周转率、应收账款周转率、固定资产周转率和自有资金周转率,分别给定各自的分数比重,通过与标准比率(行业平均比率)进行比较,确定各项指标的得分及总体指标的累计分数,从而得出企业财务状况的综合评价,继而确定其信用等级。

2. 多变量信用风险判别模型

多变量信用风险判别模型是以财务会计信息为基础,以特征财务比率为解释变量,运用数量统计方法建模。多变量信用风险判别模型主要包括以下几种。

(1) 多元线性判定模型(Z-Score 模型)

其是财务失败预警模型,最早是由 Altman (1968 年)开始研究的。该模型通过5个变量(5种财务比率)将反映企业偿债能力的指标、获利能力指标和营运能力指标有机联系起来,综合分析预测企业财务失败或破产的可能性。一般地,Z值越低,企业越有可能发生破产,具体模型为

$$Z = V_1X_1 + V_2X_2 + \dots + V_nX_n$$

式中: V_1, V_2, \dots, V_n 是权数; X_1, X_2, \dots, X_n 是各种财务比率。根据Z值的大小,可将企业分为“破产”或“非破产”两类。在实际运用时,需要将企业样本分为预测样本和测试样本,先根据预测样本构建多元线性判定模型,确定判别Z值(Z值的大小可以作为判定企业财务状况的综合标准),然后将测试样本的数据代入判别方程,得出企业的Z值,并根据判别标准进行判定。此方法还可以用于债券评级、投资决策、银行对贷款申请的评估及子公司业绩考核等。

(2) 多元逻辑模型(Logit 模型)

其采用一系列财务比率变量来预测公司破产或违约的概率,然后根据银行、投资者的风险偏好程度设定风险警界线,以此对分析对象进行风险定位和决策。Logit 模型建立在累计概率函数的基础上,不需要自变量服从多元正态分布和两组间协方差相等的条件。Logit 模型判别方法先根据多元线性判定模型确定企业破产的Z值,然后推导出企业破产的条件概率。其判别规则是,如果概率大于0.5,表明企业破产的概率比较大;如果概率低于0.5,可以判定企业为财务正常。

(3) 多元概率比回归模型 (Probit 回归模型)

其假设企业破产的概率为 p ，并假设企业样本服从标准正态分布，其概率函数的 p 位数可以用财务指标线性解释。其计算方法是先确定企业样本的极大似然函数，通过求似然函数的极大值得到参数 a 、 b ，然后利用公式，求出企业破产的概率；其判别规则与 Logit 模型判别规则相同。

(4) 联合预测模型

联合预测模型是运用企业模型来模拟企业的运作过程，动态地描述财务正常企业和财务困境企业的特征，然后根据不同特征和判别规则，对企业样本进行分类。这一模型运作的关键是准确模拟企业的运作过程，因此，它要求有一个基本的理论框架，通过这一框架来有效模拟企业的运作过程，从而能够有效反映和识别不同企业的行为特征、财务特征，并据此区分企业样本。

3. 现代金融工程模型

20 世纪 80 年代以来，受债务危机的影响，各国银行普遍重视对信用风险的管理和防范，新一代金融工程专家利用工程化的思维和数学建模技术，在传统信用风险度量的基础上提出了一系列成功的信用风险量化模型。

(1) 神经网络分析法

神经网络是从神经心理学和认识科学研究成果出发，应用数学方法发展起来的一种并行分布模式处理系统，具有高度并行计算能力、自学能力和容错能力。神经网络方法克服了传统分析过程的复杂性及选择适当模型函数形式的困难，它是一种自然的非线性建模过程，无须分清存在何种非线性关系，给建模与分析带来极大的方便。该方法用于企业财务状况研究时，一方面利用其映射能力，另一方面主要利用其泛化能力，即在经过一定数量的带噪声的样本的训练之后，网络可以抽取样本所隐含的特征关系，并对新情况下的数据进行内插和外推以推断其属性。

(2) 衍生工具信用风险的度量方法

80 年代以来，作为一种有效的避险工具，衍生工具因其在金融、投资、套期保值和利率行为中的巨大作用而获得了飞速发展。然而，这些旨在规避市场风险应运而生的衍生工具又蕴藏着新的信用风险。研究者相继提出许多方法来度量衍生工具的信用风险，最具代表性的有下列三种：一是风险敞口等值法，这种方法是以前估测信用风险敞口价值为目标，考虑了衍生工具的内在价值和时间价值，并以特殊方法处理的风险系数建立了一系列 REE 计算模型。二是模拟法，这种计算机集约型的统计方法采用蒙特卡罗模拟过程，模拟影响衍生工具价值的关键随机变量的可能路径和交易过程中各时间点或到期时的衍生工具价值，最终经过反复计算得出一个均值。三是敏感度分析法，就是利用这些比较值通过方案分析或应用风险系数来估测衍生工具价值。

(3) 集中风险的评估系统

前述方法绝大多数是度量单项贷款或投资项目的信用风险，而很少注重信用集中风险的评估。信用集中风险是所有单一项目信用风险的总和。金融机构和投资者采用贷款组合、投资组合来达到分散和化解风险的目的。1997 年，JP 摩根推出的“信用计量法”和瑞士信贷金融产品的“信用风险法”，均可以用来评估信用风险敞口亏损分布，以及计算用以

弥补风险所需的资本。“信用计量法”是以风险值为核心的动态量化风险管理系统，它集计算机技术、计量经济学、统计学和管理工程系统知识于一体，从证券组合、贷款组合的角度，全方位衡量信用风险。该方法应用的范围比较广，例如，证券、贷款、信用证、贷款承诺、衍生工具、应收账款等领域的信用风险都可用此方法进行估测。“信用风险法”是在信用评级框架下，计算每一级别或分数下的平均违约率及违约波动，并将这些因素与风险敞口综合考虑，从而算出亏损分布与所需资本预测数。

23.1.3 SPSS 中信用风险分析模块

由于测量信用风险的模型有很多种，则对应 SPSS 中的模块也有很多，例如，二元 Logistic 回归模型、分类树模型、判别分析模型，以及神经网络模型等。

二元 Logistic 回归模型就是第 9 章中我们介绍的 Logistic 回归，分类树模型为第 21 章中介绍的模型，判别分析模型为第 13 章中介绍的模型，神经网络模型在第 22 章中有所介绍，在这里不再累述，本章主要讲述信用风险中各种模型的 SPSS 应用。

下节中所用到的 SPSS 过程主要有二元 Logistic (Binary Logistic) 过程、决策树 (Tree) 过程、判别分析 (Discriminant) 过程，以及神经网络 (Neural Networks) 过程，由于神经网络过程用于信用风险分析，读者可以参看相关章节。

23.2 实例分析

23.2.1 二元 Logistic 分析过程



结果文件

——附带光盘“PROGRAM\CH23\实例 23-1”文件夹



动画演示

——附带光盘“AVI\实例 23-1.avi”文件

本实例中所使用的数据集为 SPSS 中自带的数据集 bankloan.sav，此数据集在前面章节中已经有所涉及。数据集的格式如图 23-1 所示。数据集共包含 12 个变量。

| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|----|----------|----|----|------|-------------------------------------|---|----|----|----|----|----|
| 1 | age | 数字 | 4 | 0 | Age in years | 无 | 无 | 4 | 右 | 标度 | 输入 |
| 2 | ed | 数字 | 4 | 0 | Level of education [1, Did not c... | 无 | 无 | 4 | 右 | 有序 | 输入 |
| 3 | employ | 数字 | 4 | 0 | Years with curr... | 无 | 无 | 6 | 右 | 标度 | 输入 |
| 4 | address | 数字 | 4 | 0 | Years at curren... | 无 | 无 | 7 | 右 | 标度 | 输入 |
| 5 | income | 数字 | 8 | 2 | Household inco... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 6 | debtinc | 数字 | 8 | 2 | Debt to income... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 7 | creddebt | 数字 | 8 | 2 | Credit card deb... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 8 | othdebt | 数字 | 8 | 2 | Other debt in th... | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 9 | default | 数字 | 4 | 0 | Previously defa... [0, No]... | 无 | 无 | 7 | 右 | 名义 | 输入 |
| 10 | preddef1 | 数字 | 11 | 5 | Predicted defau... | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 11 | preddef2 | 数字 | 11 | 5 | Predicted defau... | 无 | 无 | 11 | 右 | 标度 | 输入 |
| 12 | preddef3 | 数字 | 11 | 5 | Predicted defau... | 无 | 无 | 11 | 右 | 标度 | 输入 |

图 23-1 数据集的格式

下面主要是利用二元 Logistic (Binary Logistic) 过程来分析客户的信用等信息。

1. 参数设置

选择菜单“转换 (Transform) 随机数生成器 (Random Number Generators)”，然后弹出如图 23-2 所示的对话框，此对话框用于生产随机数，选择“设置起点 (Set Starting Point)”选项，激活其下的“固定值 (Fixed Value)”选项，然后在“值 (Value)”选项框中填入 9191972，单击“确定”按钮。

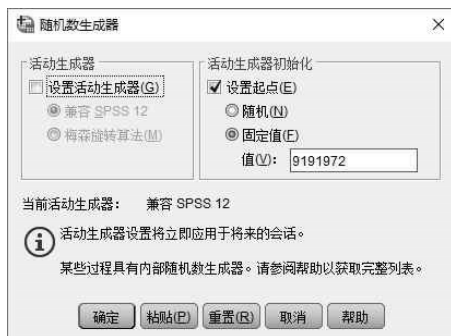


图 23-2 “随机数生成器 (Random Number Generators) 设置”对话框

选择菜单“转换 (Transform) 计算变量 (Compute Variable)”，弹出如图 23-3 所示的对话框，在“目标变量 (Target Variable)”选项栏中填入 validate，然后在“数值表达式 (Numeric Expression)”选项栏中填入 RV.BERNOULLI (0.7)。



图 23-3 “计算变量 (Compute Variable) 设置”对话框

单击图 23-3 中的“如果 (If)”按钮，弹出如图 23-4 所示的对话框，用于设置选择条件。选中“在个案满足条件时包括 (Include if Cases Statistics Condition)”选项栏，并在其下输入 MISSING(default)=0，然后单击“继续”按钮返回主界面。

设置好上述随机数产生器后，单击主界面中的“确定”按钮产生随机数。



图 23-4 “如果 (If) 设置”对话框

下面设置二元 Logistic 过程的参数, 选择菜单“分析 (Analyze) 回归 (Regression) 二元 Logistic (Binary Logistic) 过程”, 则弹出如图 23-5 所示对话框, 选择变量 default 到“因变量 (Dependent)”选项栏中, 选择变量 age、ed、employ、address、income、debtinc、creddebt、othdebt 到“协变量 (Covariates)”变量框中, 选择变量 validate 到“选择变量 (Selection Variable)”选项栏中。“方法 (Method)”选项栏的下拉菜单中选择“向前: LR (Forward: LR)”选项。

单击图 23-5 中的“规则 (Rules)”按钮, 弹出如图 23-6 所示的对话框, 在对话框中输入 1, 然后单击“继续”按钮返回主界面中。

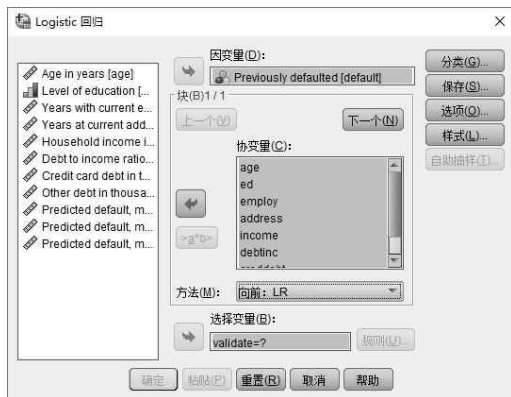


图 23-5 “二元 Logistic (Binary Logistic)”设置对话框



图 23-6 “规则 (Rules) 设置”对话框

单击图 23-5 中的“分类 (Categorical)”按钮, 弹出如图 23-7 所示的对话框, 选择变量 ed 到“分类变量 (Categorical Covariates)”选项栏中, 然后单击“继续”按钮返回主界面。

单击图 23-5 中的“保存 (Save)”按钮, 弹出如图 23-8 所示的对话框, 选择“概率 (Probabilities)”、“库克距离”、“杠杆值 (Leverage)”、“学生化 (Studentized)”选项栏, 然后单击“继续”按钮返回主界面中。

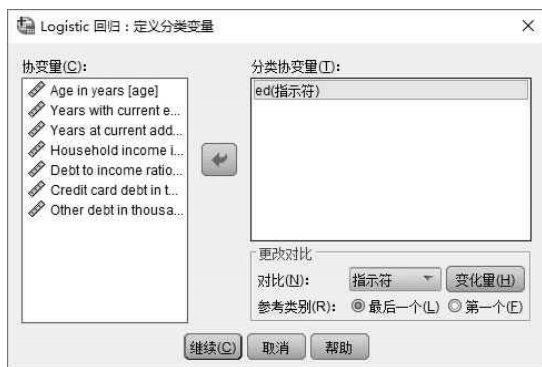


图 23-7 “分类 (Categorical) 设置”对话框



图 23-8 “保存 (Save) 设置”对话框

单击图 23-5 中的“选项 (Options)”按钮，弹出如图 23-9 所示的对话框，设置结果如图 23-9 所示，然后单击“继续”按钮返回主界面。



图 23-9 “选项 (Options) 设置”对话框

2. 结果分析

设置完成以后，单击主界面中的“确定”按钮进行分析，结果如下。图 23-10 是案例处理的结果，包括样本个数，缺失值等信息。

| 个案处理摘要 | | | |
|---------------------|------------|-----|-------|
| 未加权个案数 ^a | 个案数 | 百分比 | |
| 选定的个案 | 包括在分析中的个案数 | 499 | 58.7 |
| | 缺失个案数 | 0 | .0 |
| | 总计 | 499 | 58.7 |
| 未选定的个案 | | 351 | 41.3 |
| 总计 | | 850 | 100.0 |

^a 如果权重处于生效状态，请参阅分类表以了解个案总数。

图 23-10 案例处理的结果

然后是模型概述，给出了模型的基本信息，如图 23-11 所示，输出了考克斯-斯奈尔 R 方

统计量和内戈尔科 R 方统计量。

接着输出的是 Hosmer and Lemeshow Test 检验结果,如图 23-12 所示,从统计量检验结果可以看出第 4 步的卡方统计量为 4.027,自由度为 8,其对应的 Sig.值为 0.855,所以接受原假设,故模型的拟合结果效果比较好。

| 模型摘要 | | | |
|------|---------|-------------|----------|
| 步骤 | -2 对数似然 | 考克斯-斯泰尔 R 方 | 内戈尔科 R 方 |
| 1 | 498.012 | .116 | .172 |
| 2 | 447.301 | .201 | .299 |
| 3 | 411.553 | .257 | .381 |
| 4 | 394.721 | .281 | .417 |

图 23-11 模型输出信息

| 霍斯默-莱梅肖检验 | | | |
|-----------|--------|-----|------|
| 步骤 | 卡方 | 自由度 | 显著性 |
| 1 | 3.292 | 8 | .915 |
| 2 | 11.866 | 8 | .157 |
| 3 | 9.447 | 8 | .306 |
| 4 | 4.027 | 8 | .855 |

图 23-12 Hosmer and Lemeshow Test 检验结果

图 23-13 输出的是判别分类结果,图中给出了每步中的判别分类的正确率。从图中可以看出第 4 步中的回判正确率为 82%,说明模型的预测效果比较好。

| 分类表 ^a | | | | | | | | |
|------------------|----------------------|----------------------|-----|-------|------|----------------------|-----|-------|
| 实测 | | 选定的个案 ^b | | | | 预测 | | |
| | | Previously defaulted | | 正确百分比 | | Previously defaulted | | 正确百分比 |
| | | No | Yes | | | No | Yes | |
| 步骤 1 | Previously defaulted | No | 361 | 14 | 96.3 | 137 | 5 | 96.5 |
| | | Yes | 100 | 24 | 19.4 | 45 | 14 | 23.7 |
| | 总体百分比 | | | | 77.2 | | | 75.1 |
| 步骤 2 | Previously defaulted | No | 351 | 24 | 93.6 | 136 | 6 | 95.8 |
| | | Yes | 80 | 44 | 35.5 | 36 | 23 | 39.0 |
| | 总体百分比 | | | | 79.2 | | | 79.1 |
| 步骤 3 | Previously defaulted | No | 348 | 27 | 92.8 | 135 | 7 | 95.1 |
| | | Yes | 72 | 52 | 41.9 | 28 | 31 | 52.5 |
| | 总体百分比 | | | | 80.2 | | | 82.6 |
| 步骤 4 | Previously defaulted | No | 352 | 23 | 93.9 | 130 | 12 | 91.5 |
| | | Yes | 67 | 57 | 46.0 | 27 | 32 | 54.2 |
| | 总体百分比 | | | | 82.0 | | | 80.6 |

a. 分界值为 .500

b. 选定的个案 validate EQ 1

c. 未选定的个案 validate NE 1

d. 由于自变量中缺少值或者分类变量的值超出选定个案的范围,因此未对某些未选定的个案进行分类。

23-13 分类结果

接着输出的是模型中的方程变量信息，如图 23-14 所示，从图中可以得到二元 Logistic 模型的方程系数。图中的方框中即为方程中的系数。

| | | B | 标准误差 | 瓦尔德 | 自由度 | 显著性 | Exp(B) |
|-------------------|-------------------------------|--------|------|---------|-----|------|--------|
| 步骤 1 ^a | Debt to Income ratio (x100) | .121 | .017 | 52.676 | 1 | .000 | 1.129 |
| | 常量 | -2.476 | .230 | 116.315 | 1 | .000 | .084 |
| 步骤 2 ^b | Years with current employer | -.140 | .023 | 38.158 | 1 | .000 | .869 |
| | Debt to Income ratio (x100) | .134 | .018 | 54.659 | 1 | .000 | 1.143 |
| | 常量 | -1.621 | .259 | 39.038 | 1 | .000 | .198 |
| 步骤 3 ^c | Years with current employer | -.244 | .033 | 54.676 | 1 | .000 | .783 |
| | Debt to Income ratio (x100) | .069 | .022 | 9.809 | 1 | .002 | 1.072 |
| | Credit card debt in thousands | .506 | .101 | 25.127 | 1 | .000 | 1.658 |
| | 常量 | -1.058 | .280 | 14.249 | 1 | .000 | .347 |
| 步骤 4 ^d | Years with current employer | -.247 | .034 | 51.826 | 1 | .000 | .781 |
| | Years at current address | -.089 | .023 | 15.109 | 1 | .000 | .915 |
| | Debt to Income ratio (x100) | .072 | .023 | 10.040 | 1 | .002 | 1.074 |
| | Credit card debt in thousands | .602 | .111 | 29.606 | 1 | .000 | 1.826 |
| | 常量 | -.605 | .301 | 4.034 | 1 | .045 | .546 |

a. 在步骤 1 输入的变量: Debt to Income ratio (x100).

b. 在步骤 2 输入的变量: Years with current employer.

c. 在步骤 3 输入的变量: Credit card debt in thousands.

d. 在步骤 4 输入的变量: Years at current address.

图 23-14 方程变量信息

图 23-15 中输出的是预测概率直方图，横轴是对拖欠贷款概率的预测概率值，纵轴是观测的频数。其中的 Y 表示拖欠贷款，N 表示不拖欠贷款，从图中可以看出大部分观测都集中在 0.5 的侧面，总体看模型的拟合效果比较好。

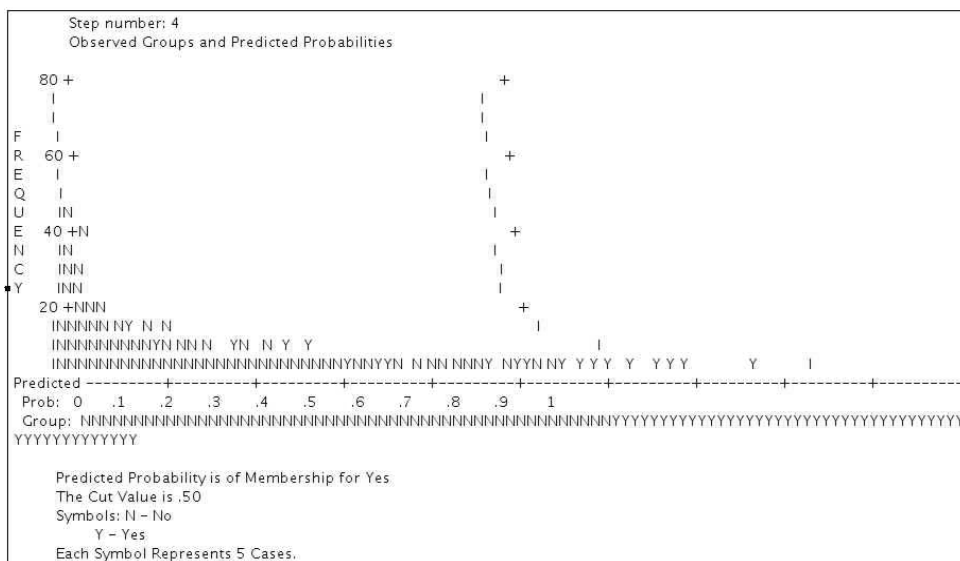
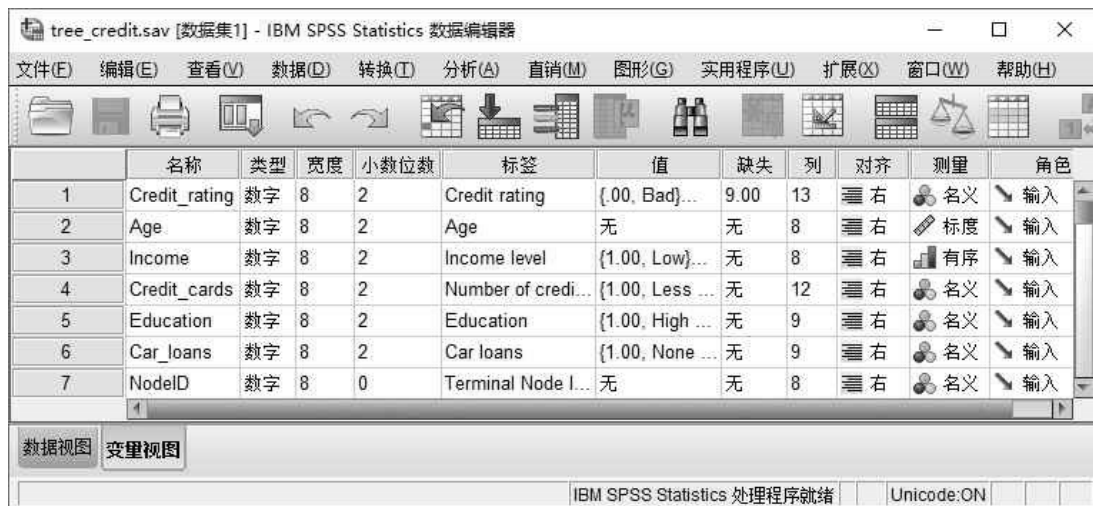


图 23-15 预测概率直方图

23.2.2 决策树分析过程

本实例主要利用 Tree 过程来研究信用风险,所用数据集为 SPSS 自带的数据文件 tree_credit.sav。使用分类树技术,银行方面可以分析及时还贷和有拖欠行为的客户特征,并能建立模型预测后续的贷款申请者拖欠银行贷款的可能性。在第 21 章中已经有所介绍。本节将继续介绍一下。

数据文件 tree_credit.sav 的数据格式如图 23-16 所示。



| | 名称 | 类型 | 宽度 | 小数位数 | 标签 | 值 | 缺失 | 列 | 对齐 | 测量 | 角色 |
|---|---------------|----|----|------|--------------------|-----------------|------|----|----|----|----|
| 1 | Credit_rating | 数字 | 8 | 2 | Credit rating | {.00, Bad}... | 9.00 | 13 | 右 | 名义 | 输入 |
| 2 | Age | 数字 | 8 | 2 | Age | 无 | 无 | 8 | 右 | 标度 | 输入 |
| 3 | Income | 数字 | 8 | 2 | Income level | {1.00, Low}... | 无 | 8 | 右 | 有序 | 输入 |
| 4 | Credit_cards | 数字 | 8 | 2 | Number of credi... | {1.00, Less ... | 无 | 12 | 右 | 名义 | 输入 |
| 5 | Education | 数字 | 8 | 2 | Education | {1.00, High ... | 无 | 9 | 右 | 名义 | 输入 |
| 6 | Car_loans | 数字 | 8 | 2 | Car loans | {1.00, None ... | 无 | 9 | 右 | 名义 | 输入 |
| 7 | NodeID | 数字 | 8 | 0 | Terminal Node I... | 无 | 无 | 8 | 右 | 名义 | 输入 |

图 23-16 数据集 tree_credit.sav 的数据格式

从第 21 章的介绍中可以看出,模型的正确分类率只有一个低于 80%。这可以从大多数端点中反映出来,预测分类(节点中反显的分类)与实际分类相同,正确率为 80%或更高。下面就对原始模型进行精练。

第 21 章的输出树形图如图 23-17 所示。

在图 23-17 中有一个端点,个案在良好信用等级与不良信用等级间几乎是平分的。在节点 9 中,预测信用等级是“良好”,但是该节点实际只有 56%的个案有良好信用等级。这就意味着在节点 9 有几乎 1/2 的个案(44%)有错误的预测分类。如果主要关心的是识别不良信用等级,这个节点不能令人满意。

1. 在节点中选择个案

考察一下节点 9 中的个案看看数据是否揭示一些有用的附加信息。在 Viewer 中双击树模型打开树编辑窗口。选择节点 9(如果想要选择多个节点,使用 Ctrl+click)。

然后从树编辑菜单中选择菜单,如图 23-18 所示。

选择上述菜单后,弹出如图 23-19 所示的对话框。

筛选个案对话框用于建立筛选变量和基于所选变量的值设置一个筛选。预设的筛选变量名是 filter_9。

- 来自所选节点的个案筛选后赋予值 1。
- 所有其他的个案赋予值 0 并在后续的分析中被排除直到改变筛选条件为止。

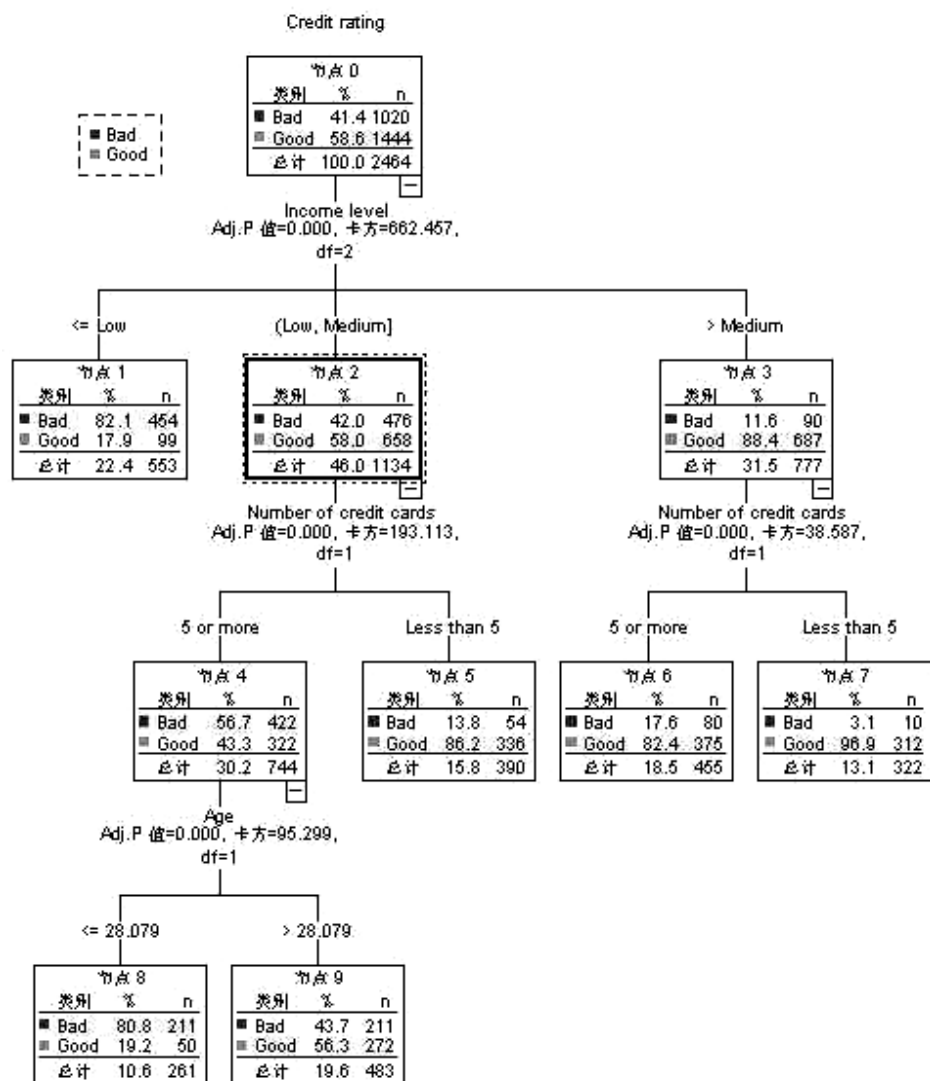


图 23-17 树形图



图 23-18 树编辑菜单



图 23-19 Fitter Cases 对话框

本例中，从现在开始不是节点 9 的个案将被筛选掉（不是删除）。要建立筛选变量并应用筛选条件，单击“确定”按钮。如图 23-20 所示。在数据编辑窗口中，被筛选掉的个案在行数上有一个斜对角线。没有标示节点 9 的个案已经筛选掉了。

| | Predicted Probabilit y_1 | Predicted Probabilit y_2 | NodeID_1 | Predicted Value_1 | Predicted Probabilit y_1_1 | Predicted Probabilit y_2_1 | filter_9 |
|----|-----------------------------|-----------------------------|----------|-------------------|-------------------------------|-------------------------------|----------|
| 1 | .44 | .56 | 9 | 1.00 | .44 | .56 | 1.00 |
| 2 | .81 | .19 | 8 | .00 | .81 | .19 | .00 |
| 3 | .82 | .18 | 1 | .00 | .82 | .18 | .00 |
| 4 | .82 | .18 | 1 | .00 | .82 | .18 | .00 |
| 5 | .44 | .56 | 9 | 1.00 | .44 | .56 | 1.00 |
| 6 | .44 | .56 | 9 | 1.00 | .44 | .56 | 1.00 |
| 7 | .44 | .56 | 9 | 1.00 | .44 | .56 | 1.00 |
| 8 | .82 | .18 | 1 | .00 | .82 | .18 | .00 |
| 9 | .82 | .18 | 1 | .00 | .82 | .18 | .00 |
| 10 | .81 | .19 | 8 | .00 | .81 | .19 | .00 |

图 23-20 筛选变量结果

从图中可以看出节点 9 的个案没有被筛选掉，所以后续的分析只包括来自节点 9 的个案。

2. 检查选择的个案

作为检查节点 9 中个案的第一步，也许想看看模型中有没有使用的变量。在这个例子中，数据文件中的所有变量包含在分析中，但是有两个变量不在最终模型中：教育和汽车贷款。

为什么程序将它们从最终模型中去掉？也许有好的理由，模型不可能告诉我们很多，但是无论如何要看一看究竟。

选择菜单“分析（Analyze） 描述统计（Descriptive Statistics） 交叉表（Crosstabs）”，然后弹出如图 23-21 所示对话框。选择信用等级作为行变量，教育和汽车贷款作为列变量。

单击图 23-21 中的“单元格（Cells）”按钮，弹出如图 23-22 所示的对话框。在“百分比”组选择“行”选项。然后单击“继续”按钮，在交叉表对话框中单击“确定”按钮运行该过程。



图 23-21 “交叉表（Crosstabs）设置”对话框

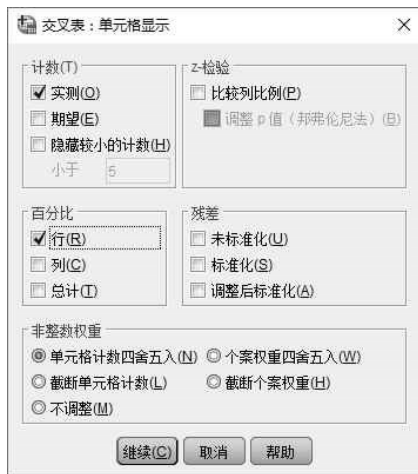


图 23-22 “单元格（Cells）设置”对话框

检查交叉表，可以看出不在模型中的两个变量，在良好信用等级和不良信用等级间的个案数没有太大的差别。

运行后的结果首先是如图 23-23 所示的对于教育变量的列联表，如图 23-24 所示的是对于汽车贷款变量的列联表。

| Credit rating * Education 交叉表 | | | | | |
|-------------------------------|----------------------|-------------|---------|--------|--|
| | | Education | | | |
| | | High school | College | 总计 | |
| Credit rating: Bad | 计数 | 110 | 101 | 211 | |
| | 占 Credit rating 的百分比 | 52.1% | 47.9% | 100.0% | |
| Good | 计数 | 128 | 144 | 272 | |
| | 占 Credit rating 的百分比 | 47.1% | 52.9% | 100.0% | |
| 总计 | 计数 | 238 | 245 | 483 | |
| | 占 Credit rating 的百分比 | 49.3% | 50.7% | 100.0% | |

图 23-23 教育变量的列联表

| Credit rating * Car loans 交叉表 | | | | | |
|-------------------------------|----------------------|-----------|-------------|--------|--|
| | | Car loans | | | |
| | | None or 1 | More than 2 | 总计 | |
| Credit rating: Bad | 计数 | 18 | 193 | 211 | |
| | 占 Credit rating 的百分比 | 8.5% | 91.5% | 100.0% | |
| Good | 计数 | 39 | 233 | 272 | |
| | 占 Credit rating 的百分比 | 14.3% | 85.7% | 100.0% | |
| 总计 | 计数 | 57 | 426 | 483 | |
| | 占 Credit rating 的百分比 | 11.8% | 88.2% | 100.0% | |

图 23-24 汽车贷款变量的列联表

对教育变量，不良信用等级组受过高中教育的个案数比受过大学教育的个案数稍微多一点，110 对 101。良好信用等级组受过大学教育的个案数比受过高中教育的个案数也多了，为 144 对 128——但是这种差别没有统计意义。

对汽车贷款变量，只有一辆车或没有汽车贷款的良好信用等级的百分比高于其对应的不良信用等级的百分比，为 14.4%对 8.5%。但是在有两辆车贷款的组中，两者的比例相差不多，为 91.5%对 85.7%。

3. 给结果分配成本

如前面的解释，除了节点 9 中个案在两个信用等级分类组中几乎平分的事实外，如果主要的目的是建立正确地识别不良信用风险的模型，预测分类是“良好”的事实就值得怀疑。如图 23-25 所示的节点 9。

虽然不能够改变节点 9 的性质，还是可以精炼模型改进不良信用等级个案的正确分类率——虽然这样也会导致对良好信用等级个案的错分率的提高。

首先，需要去掉对个案的筛选以便再次分析时所有个案都参与。

选择菜单“数据 (Data) 选择个案 (Select Cases)”，则弹出如图 23-26 所示的对话框，在“选择个案对话框”中选择全部个案 (All Cases) 选项，然后单击“确定 (OK)”按钮。

> 28.079

节点 9

| 类别 | % | n |
|------|------|-----|
| Bad | 43.7 | 211 |
| Good | 56.3 | 272 |
| 总计 | 19.6 | 483 |

图 23-25 节点 9 的信息



图 23-26 “选择个案 (Select Cases) 设置”对话框

最后打开分类树对话框，单击选项 (Options) 按钮，如图 23-27 所示。单击“错误分类成本 (the Misclassification Costs)”标签。选择“定制 (Custom)”选项，在实际分类为 Bad 和预测分类为 Good 间输入值 2。这样做告诉程序将不良信用风险作为良好信用风险的错分“成本”是将良好信用风险当做不良信用风险的错分“成本”的两倍。



图 23-27 “选项 (Options) 设置”对话框

单击“继续”按钮，然后在主对话框中单击“确定”按钮运行该过程。初看起来，由经过上述处理产生的树模型与原始树模型一样，然而，仔细的检查发现虽然每个节点的个案分布没有改变，但部分预测分类已经改变。

对端点来讲，预测分类同前面的没有区别除了节点 9 之外。预测分类现在是“不良”，即便在“良好”分类组中个案数多于 50%。因为我们告诉程序错分将不良信用风险当做良好信用风险有较高的成本，现在两个分类中个案数基本平分的任何节点有一个“不良”的预测分类，即便在“良好”分类中良好的个案数比不良的个案数多。

预测分类中这种改变反映在分类表中。

如图 23-28 所示为判别分类结果。现在不良信用等级被正确划分的比例是 85.9%，而前面是 65%。另一方面，良好信用等级的正确分类从 90%降到 70.8%，总正确分类率从 79.5%降到 77.1%。

| 实测 | 分类 | | 正确百分比 |
|--------------------|-------|-------|-------|
| | Bad | Good | |
| Bad | 876 | 144 | 85.9% |
| Good | 421 | 1023 | 70.8% |
| 总体百分比 | 52.6% | 47.4% | 77.1% |
| 生长法: CHAID | | | |
| 因变量: Credit rating | | | |

图 23-28 判别分类结果

注意：风险估计和总正确分类率彼此不再一致。你期望的风险估计是 0.229，如果总正确分类的百分比是 77.1%。在这个例子中，增加不良信用错分成本放大了风险值，使它的解释变得不直截了当了。

4. 总结

我们可以使用树模型按照某种特征将个案划分到已识别的组中，例如，银行按照客户贷款记录信用程度的特征来划分客户。如果一个详细的预测结果比其他的结果重要，可以通过给那个结果赋予较高的错分成本来精练模型——但是对一个结果降低错分率将导致另一个结果的错分率增加。

23.2.3 判别式分析过程

本实例中所使用的数据集为 SPSS 中自带的数据集 bankloan.sav，此数据集在前面已经有所涉及。数据集的格式如图 23-1 所示。数据集共包含 12 个变量。首先从数据集 850 个观测中提取出 700 个观测样本数据进行创建判别分析模型，然后利用所创建的模型来对余下的 150 个观测样本进行判别分析。

首先是设置随机数的产生器，在 23.2.1 节中已经介绍过了，在此示例中不再介绍，用户可以参考上面的 23.2.1 节中的设置情况。

然后选择菜单“分析 (Analyze) 分类 (Classify) 判别式 (Discriminant)”，弹出如图 23-29 所示的对话框，选择变量 Previously defaulted 到“分组变量 (Grouping Variable)”选项栏中，选择变量 Years with current employer、Years at current address、Debt to income ratio，以及 Credit card debt 到“自变量 (Independent)”选项栏中，选择变量 validate 到“选择变量框 (Selection Variable)”选项栏中。

然后单击图 23-29 中的“定义范围 (Define Range)”按钮，则弹出如图 23-30 所示的对话框，在“最小值 (Minimum)”选项栏中填入 0，在“最大值 (Maximum)”选项栏中填入 1，然后单击“继续”按钮返回主界面。

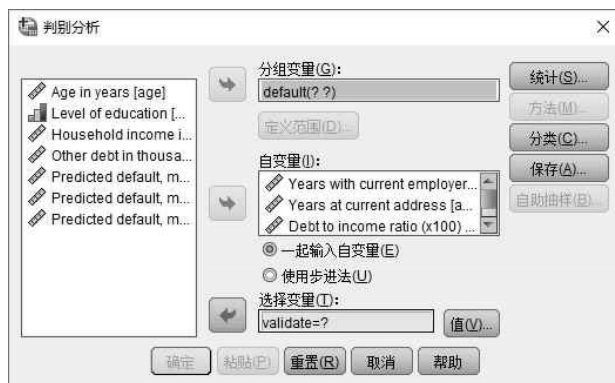


图 23-29 “判别分析 (Discriminant) 设置”对话框

单击图 23-29 中的“值 (Value)”按钮，弹出如图 23-31 所示的对话框，在“选择变量的值 (Value for Selection Variable)”选项栏中填入 1，然后单击“继续”按钮返回主界面。



图 23-30 “定义范围 Define Range”设置对话框 图 23-31 “值 (Value) 设置对话框

单击“统计量 (Statistics)”按钮，弹出如图 23-32 所示的对话框，选择“平均值 (Means)、单变量 ANOVA (Univariate ANOVAs)”、“博克斯 M (Box's M)”、“费希尔 (Fisher)”、“未标准化 (Unstandardized)”，以及“组内相关性 (Within-groups Correlation)”选项栏，然后单击“继续”按钮返回主界面。

接着单击“分类 (Classify)”按钮，弹出如图 23-33 所示的对话框，选择“摘要表 (Summary Table)”选项栏，以及“留一分类 (Leave-one-out Classification)”选项栏，然后单击“继续”按钮返回主界面。



图 23-32 “统计量 (Statistics) 设置”对话框

图 23-33 “分类 (Classify) 设置”对话框

单击主界面中的“保存 (Save)”按钮,弹出如图 23-34 所示的对话框,选择“预测组成员 (Predicted Group Membership)”选项栏,以及“组成员概率 (Probabilities of Group Membership)”选项栏,然后单击“继续”按钮返回主界面。



图 23-34 “保存 (Save) 设置”对话框

2. 结果分析

设置完成以后单击主界面“判别分析”对话框 (Discriminant Analysis Dialog Box) 中的“确定”按钮进行判别分析,首先是如图 23-35 所示的案例分析结果摘要。

| 分析个案处理摘要 | | |
|-----------------------------|-----|-------|
| 未加权个案数 | 个案数 | 百分比 |
| 有效 | 488 | 57.4 |
| 排除 | | |
| 缺失或超出范围组代码 | 150 | 17.6 |
| 至少一个缺失判别变量 | 0 | .0 |
| 既包括缺失或超出范围组代码,也包括至少一个缺失判别变量 | 0 | .0 |
| 未选中 | 212 | 24.9 |
| 总计 | 362 | 42.6 |
| 总计 | 850 | 100.0 |

图 23-35 案例分析结果信息

然后查看费希尔判别函数的系数,如图 23-36 所示,从图中可以得到没有拖欠贷款和拖欠贷款的判别函数。

| 分类函数系数 | | |
|-------------------------------|----------------------|--------|
| | Previously defaulted | |
| | No | Yes |
| Years with current employer | .270 | .097 |
| Years at current address | .150 | .112 |
| Debt to income ratio (x100) | .274 | .369 |
| Credit card debt in thousands | -.628 | -.188 |
| (常量) | -3.452 | -3.852 |
| 费希尔线性判别函数 | | |

图 23-36 费希尔判别函数

图 23-37 输出的是组内协方差矩阵,从图中可以看出,最大的相关系数发生在变量 Credit card debt in thousands 和其他变量之间。

| 汇聚组内矩阵 | | | | | |
|--------|-------------------------------|-----------------------------|--------------------------|-----------------------------|-------------------------------|
| | | Years with current employer | Years at current address | Debt to income ratio (x100) | Credit card debt in thousands |
| 相关性 | Years with current employer | 1.000 | .333 | .083 | .536 |
| | Years at current address | .333 | 1.000 | .060 | .277 |
| | Debt to income ratio (x100) | .083 | .060 | 1.000 | .440 |
| | Credit card debt in thousands | .536 | .277 | .440 | 1.000 |

图 23-37 组内协方差矩阵

图 23-38 是博克斯 M 检验结果,从图中可以看出,显著性的值为 0.000 小于 0.10,故拒绝原假设,所以应该使用分组的协方差矩阵来分析。

| 检验结果 | | |
|--------------------|-------|------------|
| 博克斯 M | | 267.083 |
| F | 近似 | 26.397 |
| | 自由度 1 | 10 |
| | 自由度 2 | 326598.755 |
| | 显著性 | .000 |
| 对等群体协方差矩阵的原假设进行检验。 | | |

图 23-38 博克斯 M 检验结果

图 23-39 输出的是威尔克 Lambda 检验结果,对应的显著性值均为 0.000 小于 0.10。

| 组平均值的同等检验 | | | | | |
|-------------------------------|------------|--------|-------|-------|------|
| | 威尔克 Lambda | F | 自由度 1 | 自由度 2 | 显著性 |
| Years with current employer | .926 | 38.747 | 1 | 486 | .000 |
| Years at current address | .980 | 10.053 | 1 | 486 | .002 |
| Debt to income ratio (x100) | .846 | 88.354 | 1 | 486 | .000 |
| Credit card debt in thousands | .933 | 34.958 | 1 | 486 | .000 |

图 23-39 组均值的均等性检验结果

接着输出的是特征值,如图 23-40 所示,特征值给出了典型判别函数中各个变量的标准化系数,由此可以判断函数主要受哪些变量的影响。

| 特征值 | | | | |
|------------------------|-------------------|-------|-------|-------|
| 函数 | 特征值 | 方差百分比 | 累计百分比 | 典型相关性 |
| 1 | .399 ^a | 100.0 | 100.0 | .534 |
| a. 在分析中使用了前 1 个典型判别函数。 | | | | |

图 23-40 特征值

最后输出的是判别分析结果，如图 23-41 所示，表中给出了经过判别函数判断后的结果，即各种回判结果的正确率。

| 分类结果 ^{a,b,d} | | | | | | |
|-----------------------|-------------------|----|----------------------|------|------|-------|
| | | | 预测组成员信息 | | | |
| | | | Previously defaulted | No | Yes | 总计 |
| 选中个案 | 原始 | 计数 | No | 263 | 87 | 350 |
| | | | Yes | 36 | 102 | 138 |
| | | | 未分组个案 | 68 | 33 | 101 |
| | | % | No | 75.1 | 24.9 | 100.0 |
| | | | Yes | 26.1 | 73.9 | 100.0 |
| | | | 未分组个案 | 67.3 | 32.7 | 100.0 |
| | 交叉验证 ^c | 计数 | No | 260 | 90 | 350 |
| | | | Yes | 37 | 101 | 138 |
| | | | 未分组个案 | 67.3 | 32.7 | 100.0 |
| | | % | No | 74.3 | 25.7 | 100.0 |
| | | | Yes | 26.8 | 73.2 | 100.0 |
| | | | 未分组个案 | 59.2 | 40.8 | 100.0 |
| 未选中个案 | 原始 | 计数 | No | 136 | 31 | 167 |
| | | | Yes | 14 | 31 | 45 |
| | | | 未分组个案 | 29 | 20 | 49 |
| | | % | No | 81.4 | 18.6 | 100.0 |
| | | | Yes | 31.1 | 68.9 | 100.0 |
| | | | 未分组个案 | 59.2 | 40.8 | 100.0 |

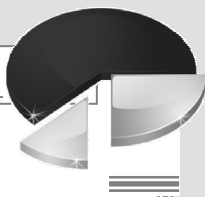
a. 正确地对 74.8% 个选定的原始已分组个案进行了分类。

b. 正确地对 78.8% 个未选定的原始已分组个案进行了分类。

c. 仅针对分析中的个案进行交叉验证。在交叉验证中，每个个案都由那些从该个案以外的所有个案派生的函数进行分类。

d. 正确地对 74.0% 个选定的进行了交叉验证的已分组个案进行了分类。

图 23-41 判别分析结果



第 24 章 SPSS 在社会经济综合评价中的应用

SPSS 在社会经济综合评价中有着广泛的应用,包括财政金融、市场研究、邮电通信、制造业、零售业、电子商务、交通运输、医疗卫生、教育机构,以及政府机构等。

本章将讲述 SPSS 在社会经济综合评价中的各种应用,如在区域经济研究,宏观经济分析等研究案例。



本讲内容

- 沿海省市经济综合指标的主成分分析
- 中国城镇居民消费结构的聚类分析研究
- 我国内地可支配收入和消费性支出之间的回归分析

24.1 沿海省市经济综合指标的主成分分析

主成分分析和因子分析在社会经济统计综合评价中是两个常被使用的统计分析方法。下面将通过 SPSS 主成分分析来研究沿海省市经济综合指标。

主成分分析中的数学模型、步骤,以及 SPSS 中主成分分析操作在前面的章节中已经介绍过,本节将不再累述。



结果文件——附带光盘“PROGRAM\实例 24-1”文件夹



动画演示——附带光盘“AVI\实例 24-1.avi”文件

1. 指标选取

选取的数据来自国家统计局网站中的统计数据,在沿海 10 省市经济状况主要指标体系中选取了 10 个指标,即

- X_1 ——GDP ;
- X_2 ——人均 GDP ;

- X_3 ——农业增加值；
- X_4 ——工业增加值；
- X_5 ——第三产业增加值；
- X_6 ——固定资产投资；
- X_7 ——基本建设投资；
- X_8 ——社会消费品零售总额；
- X_9 ——海关出口总额；
- X_{10} ——地方财政收入。

表 24-1 给出了本实例所需的数据集，为沿海省市的经济数据。

表 24-1 沿海省市的经济数据

单位：亿元

| 地区 | GDP | 人均 GDP | 农业增加值 | 工业增加值 | 第三产业增加值 | 固定资产投资 | 基本建设投资 | 社会消费品零售总额 | 海关出口总额 | 地方财政收入 |
|----|--------|--------|---------|--------|---------|--------|--------|-----------|--------|--------|
| 辽宁 | 5458.2 | 13000 | 14883.3 | 1376.2 | 2258.4 | 1315.9 | 529 | 2258.4 | 123.7 | 399.7 |
| 山东 | 10550 | 11643 | 1390 | 3502.5 | 3851 | 2288.7 | 1070.7 | 3181.9 | 211.1 | 610.2 |
| 河北 | 6076.6 | 9047 | 950.2 | 1406.7 | 2092.6 | 1161.6 | 597.1 | 1968.3 | 45.9 | 302.3 |
| 天津 | 2022.6 | 22068 | 83.9 | 822.8 | 960 | 703.7 | 361.9 | 941.4 | 115.7 | 171.8 |
| 江苏 | 10636 | 14397 | 1122.6 | 3536.3 | 3967.2 | 2320 | 1141.3 | 3215.8 | 384.7 | 643.7 |
| 上海 | 5408.8 | 40627 | 86.2 | 2196.2 | 2755.8 | 1970.2 | 779.3 | 2035.2 | 320.5 | 709 |
| 浙江 | 7670 | 16570 | 680 | 2356.5 | 3065 | 2296.6 | 1180.6 | 2877.5 | 294.2 | 566.9 |
| 福建 | 4682 | 13510 | 663 | 1047.1 | 1859 | 964.5 | 397.9 | 1663.3 | 173.7 | 272.9 |
| 广东 | 11770 | 15030 | 1023.9 | 4224.6 | 4793.6 | 3022.9 | 1275.5 | 5013.6 | 1843.7 | 1202 |
| 广西 | 2437.2 | 5062 | 591.4 | 367 | 995.7 | 542.2 | 352.7 | 1025.5 | 15.1 | 186.7 |

2. SPSS 操作

选择菜单“分析 (Analyze) 降维 (Data Reduction) 因子分析 (Factor Analysis)”，则弹出如图 24-1 所示的对话框，把变量 GDP、人均 GDP、农业增加值、工业增加值、第三产业增加值、固定资产投资、基本建设投资、社会消费品零售总额、海关出口总额、地方财政收入选入“变量 (Variables)”选项栏中。

单击主界面中的“描述 (Descriptives)”按钮，弹出如图 24-2 所示的对话框，在“相关性矩阵 (Correlation Matrix)”选项栏中选中“系数 (Coefficients)”选项，然后单击“继续”按钮，返回“因子分析 (Factor Analysis)”对话框。



图 24-1 “因子分析 (Factor Analysis) 设置”对话框

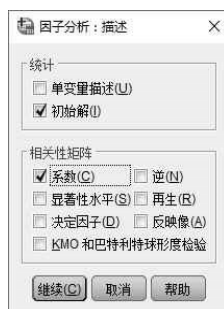


图 24-2 “描述 (Descriptives) 设置”对话框

SPSS 在调用因子分析 (Factor Analysis) 过程进行分析时, SPSS 会自动对原始数据进行标准化处理, 所以, 在得到计算结果后指的变量都是指经过标准化处理后的变量, 但 SPSS 不会直接给出标准化后的数据, 如需要得到标准化数据, 则需调用描述 (Descriptives) 过程进行计算。

设置上述参数以后就可以进行因子分析了。

3. 结果分析

单击主界面中的“确定”按钮进行分析, 结果如下。首先输出的是相关系数矩阵, 如图 24-3 所示。图中给出了 10 个变量之间的相关系数矩阵。

从图 24-3 可知 GDP 与工业增加值、第三产业增加值、固定资产投资、基本建设投资、社会消费品零售总额、地方财政收入这几个指标存在着极其显著的关系, 与海关出口总额存在着显著关系。可见许多变量之间直接的相关性比较强, 证明它们存在信息上的重叠。

| 相关性矩阵 | | | | | | | | | | | |
|-------|-----------|-------|--------|-------|-------|---------|--------|--------|-----------|--------|--------|
| | | GDP | 人均 GDP | 农业增加值 | 工业增加值 | 第三产业增加值 | 固定资产投资 | 基本建设投资 | 社会消费品零售总额 | 海关出口总额 | 地方财政收入 |
| 相关性 | GDP | 1.000 | -.094 | -.052 | .967 | .979 | .923 | .922 | .941 | .637 | .826 |
| | 人均GDP | -.094 | 1.000 | -.171 | .113 | .074 | .214 | .093 | -.043 | .081 | .273 |
| | 农业增加值 | -.052 | -.171 | 1.000 | -.132 | -.050 | -.098 | -.176 | .013 | -.125 | -.086 |
| | 工业增加值 | .967 | .113 | -.132 | 1.000 | .985 | .963 | .939 | .935 | .705 | .898 |
| | 第三产业增加值 | .979 | .074 | -.050 | .985 | 1.000 | .973 | .940 | .962 | .714 | .913 |
| | 固定资产投资 | .923 | .214 | -.098 | .963 | .973 | 1.000 | .971 | .937 | .717 | .934 |
| | 基本建设投资 | .922 | .093 | -.176 | .939 | .940 | .971 | 1.000 | .897 | .624 | .848 |
| | 社会消费品零售总额 | .941 | -.043 | .013 | .935 | .962 | .937 | .897 | 1.000 | .836 | .929 |
| | 海关出口总额 | .637 | .081 | -.125 | .705 | .714 | .717 | .624 | .836 | 1.000 | .882 |
| | 地方财政收入 | .826 | .273 | -.086 | .898 | .913 | .934 | .848 | .929 | .882 | 1.000 |

图 24-3 相关系数矩阵

然后是方差分解主成分提取分析表, 如图 24-4 所示, 主成分个数提取原则为主成分对应的特征值大于 1 的前 m 个主成分。

注意: 特征值在某种程度上可以看作表示主成分影响力度大小的指标, 如果特征值小于 1, 说明该主成分的解力度还不如直接引入一个原变量的平均解力度大, 因此一般可以用特征值大于 1 作为纳入标准。

通过图 24-4 可知, 提取 2 个主成分, 即 $m=2$ 。如图 24-5 所示的是初始因子载荷矩阵。

| 总方差解释 | | | | | | |
|-------|------------|------------|---------|---------|-------|--------|
| 成分 | 总计 | 初始特征值 | | 提取载荷平方和 | | |
| | | 总计 | 方差百分比 | 累积 % | 总计 | 方差百分比 |
| 1 | 7.220 | 72.205 | 72.205 | 72.205 | 7.220 | 72.205 |
| 2 | 1.235 | 12.346 | 84.551 | 84.551 | 1.235 | 12.346 |
| 3 | .877 | 8.769 | 93.319 | | | |
| 4 | .547 | 5.466 | 98.786 | | | |
| 5 | .085 | .854 | 99.640 | | | |
| 6 | .021 | .211 | 99.850 | | | |
| 7 | .012 | .119 | 99.970 | | | |
| 8 | .002 | .018 | 99.988 | | | |
| 9 | .001 | .012 | 100.000 | | | |
| 10 | -1.098E-16 | -1.098E-15 | 100.000 | | | |

提取方法: 主成分分析法。

图 24-4 主成分提取分析表

从图 24-5 可知 GDP、工业增加值、第三产业增加值、固定资产投资、基本建设投资、社会消费品零售总额、海关出口总额、地方财政收入在第一主成分上有较高载荷,说明第一主成分基本反映了这些指标的信息;人均 GDP 和农业增加值指标在第二主成分上有较高载荷,说明第二主成分基本反映了人均 GDP 和农业增加值两个指标的信息。所以,提取两个主成分是可以基本反映全部指标的信息,所以,决定用两个新变量来代替原来的 10 个变量。但这两个新变量的表达还不能从输出窗口中直接得到,因为“Component Matrix”是指初始因子载荷矩阵,每一个载荷量表示主成分与对应变量的相关系数。

| 成分矩阵 ^a | | |
|-------------------|-------|-------|
| | 成分 | |
| | 1 | 2 |
| GDP | .949 | .195 |
| 人均GDP | .112 | -.824 |
| 农业增加值 | -.109 | .677 |
| 工业增加值 | .978 | -.005 |
| 第三产业增加值 | .986 | .070 |
| 固定资产投资 | .983 | -.068 |
| 基本建设投资 | .947 | -.024 |
| 社会消费品零售总额 | .977 | .176 |
| 海关出口总额 | .800 | -.051 |
| 地方财政收入 | .954 | -.128 |

提取方法: 主成分分析法。
a. 提取了 2 个成分。

图 24-5 初始因子载荷矩阵

用图 24-5 中的数据除以主成分相对应的特征值开平方根便得到两个主成分中每个指标所对应的系数, 即

$$F_1 = 0.353ZX_1 + 0.042ZX_2 - 0.041ZX_3 + 0.364ZX_4 + 0.367ZX_5 + 0.366ZX_6 + 0.352ZX_7 + 0.364ZX_8 + 0.298ZX_9 + 0.355ZX_{10}$$

$$F_2 = 0.175ZX_1 - 0.741ZX_2 + 0.609ZX_3 - 0.004ZX_4 + 0.063ZX_5 - 0.061ZX_6 - 0.022ZX_7 + 0.158ZX_8 - 0.046ZX_9 - 0.115ZX_{10}$$

以每个主成分所对应的特征值占所提取主成分总的特征值之和的比例作为权重计算主成分综合模型, 即

$$F = \frac{\lambda_1}{\lambda_1 + \lambda_2} F_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} F_2$$

可得到主成分综合模型如下, 即

$$F = 0.327ZX_1 - 0.072ZX_2 + 0.054ZX_3 + 0.310ZX_4 + 0.323ZX_5 + 0.304ZX_6 + 0.297ZX_7 + 0.334ZX_8 + 0.248ZX_9 + 0.286ZX_{10}$$

根据主成分综合模型即可计算综合主成分值, 并对其按综合主成分值进行排序, 即可对各地区进行综合评价比较, 参见表 24-2。

对得出的综合主成分(评价)值, 可用实际结果、经验与原始数据做聚类分析进行检验, 对有争议的结果, 可用原始数据做判别分析解决争议, 具体评价与检验本文不做论

述,如读者有兴趣可自行进行检验论述。

表 24-2 综合主成分值

| 城 市 | 第一主成分 F_1 | 排 名 | 第二主成分 F_2 | 排 名 | 综合主成分 F | 排 名 |
|-----|-------------|-----|-------------|-----|-----------|-----|
| 广东 | 5.23 | 1 | 0.11 | 6 | 4.48 | 1 |
| 江苏 | 2.25 | 2 | 0.23 | 5 | 1.96 | 2 |
| 山东 | 1.96 | 3 | 0.50 | 2 | 1.75 | 3 |
| 浙江 | 1.16 | 4 | -0.19 | 8 | 0.96 | 4 |
| 上海 | 0.30 | 5 | -2.36 | 10 | -0.09 | 5 |
| 辽宁 | -1.24 | 6 | 1.96 | 1 | -0.78 | 6 |
| 河北 | -1.35 | 7 | 0.41 | 4 | -1.10 | 7 |
| 福建 | -1.97 | 8 | -0.07 | 7 | -1.70 | 8 |
| 天津 | -3.04 | 9 | -1.01 | 9 | -2.74 | 9 |
| 广西 | -3.29 | 10 | 0.41 | 3 | -2.75 | 10 |

24.2 中国内地城镇居民消费结构的聚类分析

消费结构是在一定的社会经济条件下,人们(包括各种不同类型的消费者和社会集团)在消费过程中所消费的各种不同类型的消费资料(包括劳务)的比例关系。有实物和价值两种表现形式。实物形式指人们在消费中,消费了一些什么样的消费资料,以及它们各自的数量。价值形式指以货币表示的人们在消费过程中消费的各种不同类型的消费资料的比例关系。在现实生活中具体的表现为各项生活支出。目前普遍将我国经济发展状况由地域的不同分成东部地区、东北地区、中部地区和西部地区。本文利用聚类分析法对我国 31 个省(直辖市、自治区)的城镇居民消费结构进行聚类分析,以期发现我国各区域之间城镇居民消费结构的差异,从而为引导我国区域消费结构向着协调方向发展,为各地政府根据地区间消费结构差异制定更加合理的引导性政策提供更加有效的依据。

我国经济区划的分类尽管每种都包含不同类型的省份,但基本是按照地理位置进行分类的。对中国经济问题进行研究大都是以当时的经济区划为依据展开的,分析中国的消费问题也不例外。由于不同类型的省份影响其消费结构的因素不尽相同,因此,单纯地按照地理位置进行分类,以此划分为基础的进一步分析难免会产生一定的片面性。本文分类的目的是为了将消费结构相近的地区合归为一类,避免单纯按地理位置划分的不合理性,使地区分类更具代表性;也为研究中国城镇居民消费结构提供一种不同的角度。因此,本文选取构成居民消费支出的主要项目作为指标。按照中华人民共和国统计局统计口径,构成城镇居民消费性支出的项目有:食品、衣着、家庭设备用品及服务、医疗保健、交通和通信、教育文化娱乐服务、居住、杂项商品和服务,以上构成城镇居民消费性支出的 8 个项目即为所选指标。



结果文件

——附带光盘“PROGRAM\CH24\实例 24-2”文件夹



动画演示

——附带光盘“AVI\实例 24-2.avi”文件

1. 分析数据

为了消除各地区在区域面积、人口等方面的先天差异，使数据的分析结果更合理，这里的指标均采用各地区城镇居民家庭平均每人全年消费性支出作为分析对象，即采用人均值。根据中国统计年鉴，得到 2006 年的统计数据，参见表 24-3。

表 24-3 2006 年各地区城镇居民家庭平均每人全年消费性支出

单位：元

| 地 区 | 食 品 | 衣 着 | 家庭设备用品 及服务 | 医 疗 保 健 | 交通和通信 | 教育文化娱乐 服务 | 居 住 | 杂项商品和 服务 |
|-----|---------|---------|---------------|---------|---------|--------------|---------|-------------|
| 北京 | 4560.52 | 1442.42 | 977.47 | 1322.36 | 2173.26 | 2514.76 | 1212.89 | 621.74 |
| 天津 | 3680.22 | 864.89 | 634.39 | 1049.33 | 1092.87 | 1452.17 | 1368.20 | 405.99 |
| 河北 | 2492.26 | 849.58 | 460.27 | 737.43 | 875.43 | 827.72 | 864.92 | 235.88 |
| 山西 | 2252.50 | 1016.69 | 441.82 | 589.97 | 825.18 | 1007.92 | 830.38 | 206.48 |
| 内蒙古 | 2323.55 | 1168.93 | 464.55 | 555.00 | 928.48 | 1052.65 | 802.26 | 371.19 |
| 辽宁 | 3102.13 | 846.91 | 362.10 | 767.13 | 797.64 | 853.92 | 909.42 | 348.23 |
| 吉林 | 2457.21 | 907.61 | 318.65 | 671.44 | 815.02 | 890.22 | 984.95 | 307.56 |
| 黑龙江 | 2215.68 | 971.44 | 319.37 | 634.30 | 665.01 | 843.94 | 755.32 | 250.37 |
| 上海 | 5248.95 | 1026.87 | 877.59 | 762.92 | 2332.83 | 2431.74 | 1435.72 | 645.13 |
| 江苏 | 3462.66 | 886.82 | 647.52 | 600.69 | 1203.45 | 1467.36 | 997.53 | 362.56 |
| 浙江 | 4393.40 | 1383.63 | 615.45 | 852.27 | 2492.01 | 1946.15 | 1229.25 | 436.37 |
| 安徽 | 3091.28 | 869.55 | 336.99 | 441.42 | 788.25 | 869.23 | 694.17 | 203.83 |
| 福建 | 3854.26 | 784.71 | 525.65 | 513.61 | 1232.70 | 1321.33 | 1233.49 | 341.96 |
| 江西 | 2636.93 | 725.72 | 451.32 | 357.03 | 600.16 | 894.58 | 742.93 | 236.87 |
| 山东 | 2711.65 | 1091.22 | 526.29 | 624.06 | 1175.57 | 1201.97 | 838.17 | 299.48 |
| 河南 | 2215.32 | 919.31 | 431.02 | 520.57 | 762.08 | 847.12 | 737.00 | 252.76 |
| 湖北 | 2868.39 | 877.01 | 401.22 | 517.19 | 763.14 | 997.74 | 752.56 | 220.08 |
| 湖南 | 2850.94 | 868.23 | 513.63 | 632.52 | 965.09 | 1182.18 | 871.70 | 285.00 |
| 广东 | 4503.86 | 719.26 | 633.03 | 707.86 | 2394.66 | 1813.86 | 1254.69 | 405.00 |
| 广西 | 2857.40 | 477.67 | 360.62 | 401.06 | 785.01 | 850.90 | 826.86 | 232.43 |
| 海南 | 3097.71 | 375.42 | 405.81 | 369.33 | 1154.87 | 791.24 | 743.60 | 188.80 |
| 重庆 | 3415.92 | 1038.98 | 615.74 | 705.72 | 976.02 | 1449.49 | 954.56 | 242.26 |
| 四川 | 2838.22 | 754.93 | 505.83 | 449.87 | 1009.35 | 976.33 | 728.43 | 261.85 |
| 贵州 | 2649.02 | 832.74 | 446.53 | 329.77 | 775.07 | 938.37 | 627.23 | 249.66 |
| 云南 | 3102.46 | 745.08 | 335.14 | 600.08 | 1076.93 | 754.69 | 585.35 | 180.07 |
| 西藏 | 3107.90 | 734.83 | 211.10 | 221.70 | 694.21 | 359.34 | 612.67 | 250.82 |
| 陕西 | 2588.91 | 768.47 | 478.58 | 612.30 | 824.46 | 1280.14 | 746.59 | 253.84 |
| 甘肃 | 2408.37 | 854.00 | 403.80 | 562.74 | 703.07 | 1034.42 | 716.35 | 291.46 |
| 青海 | 2366.42 | 724.96 | 420.31 | 542.93 | 753.07 | 793.72 | 653.04 | 275.66 |
| 宁夏 | 2444.98 | 874.39 | 480.70 | 578.75 | 774.57 | 846.72 | 890.97 | 314.49 |
| 新疆 | 2386.97 | 953.03 | 364.11 | 472.35 | 765.72 | 819.72 | 698.66 | 269.45 |

2. 聚类分析

聚类分析是通过数据建模简化数据的一种方法。传统的统计聚类分析方法包括系统聚类法(Hierarchical Cluster Procedures)、有序样品聚类法、动态聚类法、模糊聚类法、图论聚类法、聚类预报法等。在 SPSS 中同样有对应的过程来实现聚类分析,聚类分析的具体数学模型这里不再介绍。

利用该方法进行聚类分析的主要思想和一般步骤如下。

第一步:确定基础数据,选定一种相似性度量准则,计算出相似性度量矩阵。

第二步:认为各样本自成一类,即 N 个样本就有 N 类。

第三步:将各类中最相似的两类合并为新类。

第四步:按某种求新类相似性的方法,计算新类与其余各类之间的相似性,再将其中最相似的两类合并,并重复这一步,直到最后聚成一大类为止。

选择菜单“分析(Analyze) 分类(Classify) 系统聚类(Hierarchical Cluster)”,则弹出如图 24-6 所示的对话框,选择变量地区到“标注个案(Label Cases by)”选项栏中,选择变量食品、衣着、家庭设备用品及服务、医疗保健、交通和通信、教育文化娱乐服务、居住、杂项商品和服务到“变量(Variables(s))”选项栏中。

然后单击主界面中的“图(Plots)”按钮,弹出如图 24-7 所示的对话框,选中“谱系图”选项栏,然后单击“继续”按钮返回主界面。



图 24-6 “系统聚类(Hierarchical Cluster)分析”对话框 图 24-7 “图(Plots)设置”对话框

3. 结果分析

设置完成以后单击主界面中的“确定”按钮可以进行分析,结果如下。首先是如图 24-8 所示的案例分析结果。图 24-8 中包含观察样本数等信息。

然后输出的就是聚类的迭代过程,如图 24-9 所示,给出了每一步的聚类信息,共 30 步。

最后输出的是聚类谱系图,如图 24-10 所示,从图中可以明显地看出 31 个样本数据的聚类结果。

根据谱系聚类图整理如下。

| 个案处理摘要 ^{a,b} | | | | | |
|-----------------------|-------|----------|-----|-----|-------|
| 有效 | | 个案 缺失 | | 总计 | |
| 个案数 | 百分比 | 个案数 | 百分比 | 个案数 | 百分比 |
| 31 | 100.0 | 0 | .0 | 31 | 100.0 |
| a. 平方欧氏距离 使用中 | | | | | |
| b. 平均联接 (组间) | | | | | |

图 24-8 案例分析信息汇总

| 阶段 | 组合聚类 | | 系数 | 首次出现聚类的阶段 | | 下一个阶段 |
|----|------|------|--------------|-----------|------|-------|
| | 聚类 1 | 聚类 2 | | 聚类 1 | 聚类 2 | |
| 1 | 8 | 16 | .37891.834 | 0 | 0 | 4 |
| 2 | 3 | 30 | .45839.536 | 0 | 0 | 3 |
| 3 | 3 | 7 | .52310.813 | 2 | 0 | 12 |
| 4 | 8 | 31 | .56063.788 | 1 | 0 | 9 |
| 5 | 14 | 24 | .58426.646 | 0 | 0 | 15 |
| 6 | 4 | 5 | .70556.455 | 0 | 0 | 13 |
| 7 | 12 | 17 | .80420.976 | 0 | 0 | 14 |
| 8 | 28 | 29 | .83770.426 | 0 | 0 | 9 |
| 9 | 8 | 28 | .86354.426 | 4 | 8 | 12 |
| 10 | 10 | 22 | .105740.815 | 0 | 0 | 20 |
| 11 | 18 | 23 | .111815.936 | 0 | 0 | 14 |
| 12 | 3 | 8 | .132275.172 | 3 | 9 | 13 |
| 13 | 3 | 4 | .15941.9.065 | 12 | 6 | 23 |
| 14 | 12 | 18 | .169418.965 | 7 | 11 | 16 |
| 15 | 14 | 20 | .196650.457 | 5 | 0 | 19 |
| 16 | 6 | 12 | .225647.339 | 0 | 14 | 19 |
| 17 | 21 | 25 | .227440.711 | 0 | 0 | 25 |
| 18 | 15 | 27 | .261505.899 | 0 | 0 | 21 |
| 19 | 6 | 14 | .274959.068 | 16 | 15 | 21 |
| 20 | 10 | 13 | .368181.666 | 10 | 0 | 22 |
| 21 | 6 | 15 | .379278.188 | 19 | 18 | 23 |
| 22 | 2 | 10 | .408535.324 | 0 | 20 | 29 |
| 23 | 3 | 6 | .431462.386 | 13 | 21 | 26 |
| 24 | 11 | 19 | .503361.150 | 0 | 0 | 28 |
| 25 | 21 | 26 | .537970.272 | 17 | 0 | 26 |
| 26 | 3 | 21 | .606836.753 | 23 | 25 | 29 |
| 27 | 1 | 9 | .1052122.002 | 0 | 0 | 28 |
| 28 | 1 | 11 | .1231586.783 | 27 | 24 | 30 |
| 29 | 2 | -3 | .1806363.699 | 22 | 26 | 30 |
| 30 | 1 | 2 | .8505754.249 | 28 | 29 | 0 |

图 24-9 聚类迭代信息

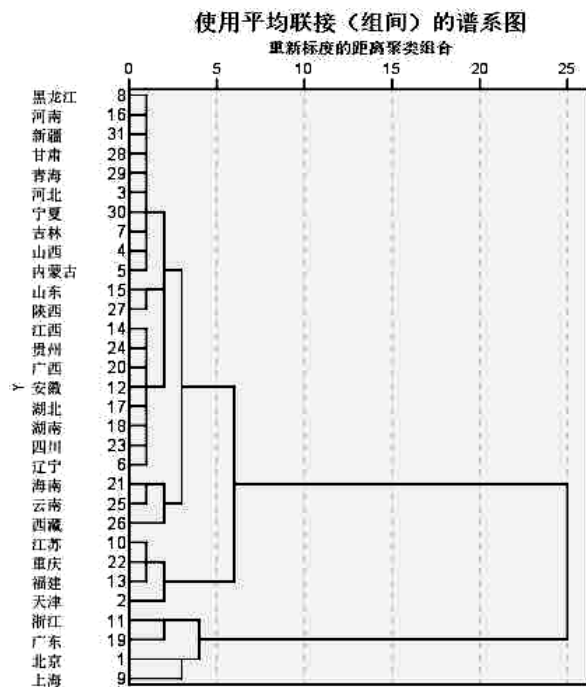


图 24-10 聚类谱系图

其中，第一类中有上海、北京、广东、浙江。第二类中包括天津、江苏、福建、重庆。第三类为海南、云南、西藏。第四类为江西、贵州、广西、安徽、湖北、湖南、四

川、辽宁。第五类包括黑龙江、河南、新疆、甘肃、青海、河北、宁夏、吉林、山西、内蒙古。第六类包括山东和陕西。

居民消费结构受地域所处的经济区域影响较大,但这也不是绝对的。例如,分类结果中同属一类的天津、江苏、福建和重庆,辽宁、安徽、江西、湖北、广西等省市,它们之间地理位置相距甚远,但同类地区的居民消费结构却相当相似。虽然从总量上来讲,地区之间经济发展水平的差距悬殊,是影响不同地域居民消费结构的重要因素,而居民消费结构的不同也会进一步影响当地经济的发展,但是图 24-10 显示的结果表明:不同地域的城镇居民的消费结构也可能相似。这种相似性的产生有着多方面的原因,例如,经济发展水平相近、居民消费观念相似、生活方式类似、生产方式近似、产业结构雷同。

24.3 我国内地可支配收入和消费性支出之间的回归分析

考察 2001 年度我国内地各省、自治区、直辖市,可支配收入(Income)和消费性支出(Expend)之间的关系,数据参见表 24-4(摘自《中国统计年鉴 2002》,单位:元)。

表 24-4 我国内地可支配收入和消费性支出

| 地 区 | 可支配收入 | 消费性支出 | 地 区 | 可支配收入 | 消费性支出 |
|-----|----------|---------|-----|----------|---------|
| 北京 | 11577.48 | 8922.72 | 湖北 | 5855.98 | 4804.79 |
| 天津 | 8958.7 | 6987.22 | 湖南 | 6780.56 | 5546.22 |
| 河北 | 5984.82 | 4479.75 | 广东 | 10415.19 | 8099.63 |
| 山西 | 5391.05 | 4123.01 | 广西 | 6665.73 | 5224.73 |
| 内蒙古 | 5535.89 | 4195.62 | 海南 | 5838.84 | 4367.85 |
| 辽宁 | 5797.01 | 4654.42 | 重庆 | 6721.09 | 5873.69 |
| 吉林 | 5340.46 | 4337.22 | 四川 | 6360.47 | 5176.17 |
| 黑龙江 | 5425.87 | 4192.36 | 贵州 | 5451.91 | 4273.9 |
| 上海 | 12883.46 | 9336.1 | 云南 | 6797.71 | 5252.6 |
| 江苏 | 7375.1 | 5532.74 | 西藏 | 7869.16 | 5994.39 |
| 浙江 | 10464.67 | 7952.39 | 陕西 | 5483.73 | 4637.74 |
| 安徽 | 5668.8 | 4517.65 | 甘肃 | 5382.91 | 4420.31 |
| 福建 | 8313.08 | 6015.11 | 青海 | 5853.72 | 4698.59 |
| 江西 | 5506.02 | 3894.51 | 宁夏 | 5544.17 | 4595.4 |
| 山东 | 7101.08 | 5252.41 | 新疆 | 6395.04 | 4931.4 |
| 河南 | 5267.42 | 4110.17 | | | |



结果文件

——附带光盘“PROGRAM\CH24\实例 24-3”文件夹



动画演示

——附带光盘“AVI\实例 24-3.avi”文件

下面可支配收入为自变量,消费性支出为因变量,试用最小二乘法确定回归方程,并就各地区可支配收入计算消费性支出的估计值。对方程的拟合情况进行诊断,解析各参数经济意义(显著性水平取 0.05)。

1. 参数设置

首先把上述数据集 CH2403 输入到 SPSS 数据窗口之中，然后选择菜单“分析 (Analyze) 回归 (Regression) 线性 (Linear)”，则弹出如图 24-11 所示的对话框。

- 选择变量消费性支出到“因变量 (Dependent)”选项栏中。
- 选择变量可支配收入到“自变量 (Independent)”选项栏中。

然后单击图 24-11 中的“统计量 (Statistics)”按钮，则弹出如图 24-12 所示的对话框，选择如下几个选项栏。

- 估算值；
- 置信区间；
- 协方差矩阵；
- 德宾—沃森；
- 模型拟合；
- R 方变化量；
- 描述。

设置完后单击“继续”按钮返回主界面。



图 24-11 “线性回归 (Linear)”对话框



图 24-12 “统计量 (Statistics)”设置对话框

单击主界面中的“图 (Plots)”按钮，弹出如图 24-13 所示的对话框，选择变量 SDRESID 和变量 ZPRED 到 Y 和 X 选项框，然后单击“下一张 (Next)”按钮，接着选择变量 ZRESID 和变量 ZPRED 到 Y 和 X 选项框。并选择“直方图 (Histogram)”选项和“正态概率图 (Normal Probability Plot)”选项，然后单击“继续”按钮返回主界面中。

接着单击主界面中的“保存 (Save)”按钮，弹出如图 24-14 所示的对话框，选择如下选项。

- 库克距离；
- 杠杆值；
- 平均值；
- 单值，并在其下的置信区间中输入 95%。

然后单击“继续”按钮返回主界面中。
对话框选项设置保持不变,选项为系统默认。



图 24-13 “图 (Plots) 设置”对话框



图 24-14 “保存 (Save) 设置”对话框

2. 结果分析

设置完上述的参数以后,则单击主界面中的“确定”按钮进行分析,结果如下,首先是模型信息输出,如图 24-15 所示。包括 R 方统计量,调整后的 R 方统计量,标准误等统计信息。

| 模型摘要 ^a | | | | | | | | | |
|-------------------|------------------|------|---------|---------|--------|----------|-------|-------|-----------|
| 模型 | R | R 方 | 调整后 R 方 | 标准估算的误差 | 更改统计 | | | | |
| | | | | | R 方变化量 | F 变化量 | 自由度 1 | 自由度 2 | 显著性 F 变化量 |
| 1 | .99 ^a | .975 | .974 | 233.818 | .975 | 1117.493 | 1 | 29 | .000 |

a. 预测变量: (常量), 可支配收入
b. 因变量: 消费性支出

图 24-15 模型信息

然后输出的是方差分析表,如图 24-16 所示,给出了回归平方和、残差平方和等信息,其中显著性等于 0.000,小于 0.05,因此,可以显著的拒绝总体回归系数为 0 的假设。

| ANOVA ^a | | | | | | |
|--------------------|--------------|-----|--------------|----------|-------------------|--|
| 模型 | 平方和 | 自由度 | 均方 | F | 显著性 | |
| 1 回归 | 61094185.970 | 1 | 61094185.970 | 1117.493 | .000 ^b | |
| 残差 | 1585451.717 | 29 | 54670.749 | | | |
| 总计 | 62679637.690 | 30 | | | | |

a. 因变量: 消费性支出
b. 预测变量: (常量), 可支配收入

图 24-16 方差分析表

图 24-17 输出的是系数有关信息，给出了所有模型的回归系数的估计值，同样可以得到回归方程，即

$$\text{消费性支出} = 427.893 + 0.716 \times \text{可支配收入}$$

| 系数 ^a | | | | | | | | |
|-----------------|--------|---------|---------|------|--------|----------------|---------|---------|
| 模型 | 未标准化系数 | | 标准化系数 | t | 显著性 | B 的 95.0% 置信区间 | | |
| | B | 标准误差 | Beta | | | 下限 | 上限 | |
| 1 | (常量) | 427.893 | 153.624 | | 2.785 | .009 | 113.697 | 742.089 |
| | 可支配收入 | .716 | .021 | .987 | 33.429 | .000 | .672 | .759 |

a. 因变量：消费性支出

图 24-17 系数估计

最后输出的是回归残差的直方图，如图 24-18 所示，并同时绘制了正态性曲线，可以看出残差基本符合正态性分布，但是也存在一个比较大的偏差，即坐标 3 出的直方。

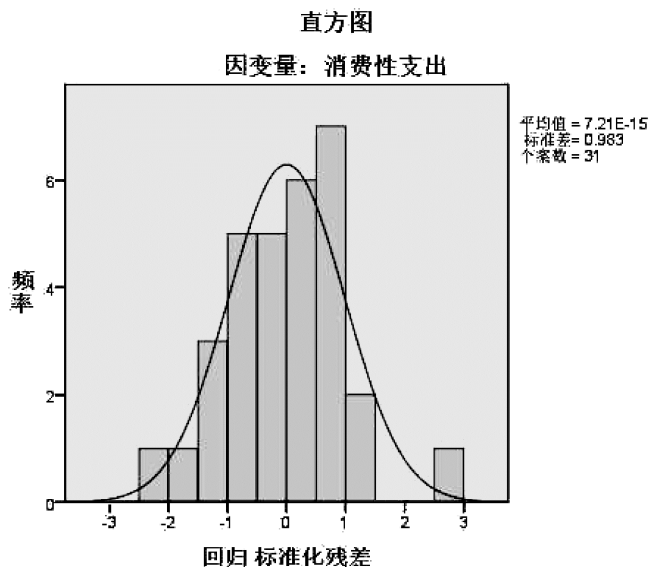


图 24-18 残差直方图